

# $\Phi$ -Entropic Measures of Correlation

Salman Beigi and Amin Gohari

School of Mathematics, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

## Abstract

A measure of correlation is said to have the *tensorization* property if it is unchanged when computed for i.i.d. copies. More precisely, a measure of correlation between two random variables  $(X, Y)$  denoted by  $\rho(X, Y)$ , has the tensorization property if  $\rho(X^n, Y^n) = \rho(X, Y)$  where  $(X^n, Y^n)$  is  $n$  i.i.d. copies of  $(X, Y)$ . Two well-known examples of such measures are the maximal correlation and the hypercontractivity ribbon (HC ribbon). We show that the maximal correlation and HC ribbons are special cases of  $\Phi$ -ribbon, defined in this paper for any function  $\Phi$  from a class of convex functions ( $\Phi$ -ribbon reduces to HC ribbon and the maximal correlation for special choices of  $\Phi$ ). Any  $\Phi$ -ribbon is shown to be a measures of correlation with the tensorization property. We show that the  $\Phi$ -ribbon also characterizes the  $\Phi$ -strong data processing inequality constant introduced by Raginsky. We further study the  $\Phi$ -ribbon for the choice of  $\Phi(t) = t^2$  and introduce an equivalent characterization of this ribbon.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Preliminaries</b>	<b>4</b>
2.1	Hypercontractivity ribbon . . . . .	4
2.2	$\Phi$ -entropy . . . . .	6
<b>3</b>	<b><math>\Phi</math>-ribbon</b>	<b>10</b>
3.1	Examples . . . . .	13
<b>4</b>	<b>Strong data processing inequalities</b>	<b>13</b>
4.1	Example: sums of i.i.d. random variables . . . . .	18
<b>5</b>	<b>Maximal correlation ribbon</b>	<b>19</b>
5.1	Alternative characterizations of the MC ribbon . . . . .	19
5.2	Extreme MC ribbons . . . . .	22
5.3	Examples . . . . .	23
5.4	Another multipartite correlation region . . . . .	25
<b>6</b>	<b>Summary of the results</b>	<b>26</b>
<b>A</b>	<b>SDPI constant</b>	<b>28</b>
<b>B</b>	<b>Proof of Theorem 14</b>	<b>30</b>
<b>C</b>	<b>Proof of Theorem 15</b>	<b>32</b>

<b>D Proof of Theorem 26</b>	<b>33</b>
<b>E Proof of Proposition 30</b>	<b>34</b>
<b>F Proof of Theorem 31</b>	<b>35</b>
<b>G Proofs of Theorem 33</b>	<b>37</b>
<b>H Proof of Theorem 35</b>	<b>38</b>

## 1 Introduction

A measure of correlation is called to have the *tensorization* property if it is unchanged when computed for i.i.d. copies. Such measures of correlations have found applications in the non-interactive distribution simulation problem [1], distributed source and channel coding problems [2], as well as simulation of non-local correlation by wirings [3]. In this paper we introduce new measures of correlation with the tensorization property that generalize two previously known such measures.

Let us explain the notion of tensorization via the example of non-interactive distribution simulation [1]. Fix some bipartite distribution  $p_{XY}$ . Suppose that two parties, Alice and Bob, are given i.i.d. samples  $X^n$  and  $Y^n$  respectively, and they are asked to output *one* sample of  $A$  and  $B$  respectively, distributed according to some predetermined distribution  $q_{AB}$ . Alice and Bob can choose  $n$  to be as large as they want, but are not allowed to communicate after receiving  $X^n$  and  $Y^n$ . The problem of deciding whether this task is feasible or not is a hard problem in general. Nevertheless, we may obtain impossibility results using the data processing inequality.

Suppose that  $I(X^n; Y^n) < I(A; B)$ . In this case, by the data processing inequality, local transformation of  $(X^n, Y^n)$  to  $(A, B)$  is infeasible. However, note that mutual information is *additive*, i.e., we have  $I(X^n; Y^n) = nI(X; Y)$ . Then, unless  $X$  and  $Y$  are independent, by choosing  $n$  to be large enough,  $I(X^n; Y^n)$  becomes as large as we want and greater than  $I(A; B)$ . Therefore, the data processing inequality of mutual information does not give us any useful bound on this problem, simply because mutual information is additive and increases when computed on i.i.d. copies. So we need to use a measure of correlation with the tensorization property as defined below.

Suppose that there is some function  $\rho(\cdot, \cdot)$  of bipartite distributions that similar to mutual information satisfies the data processing inequality (i.e., it is a measure of correlation), but instead satisfies

$$\rho(X^n, Y^n) = \rho(X, Y). \quad (1)$$

The above equation is called the *tensorization* property. Given such a measure we find that local transformation of  $(X^n, Y^n)$  to  $(A, B)$  is impossible (even for arbitrarily large  $n$ ) if  $\rho(X, Y) < \rho(A, B)$ .

**Maximal correlation:** A notable example of such a measure of correlation is *maximal correlation* [4, 5, 6, 7], which was used by Witsenhausen [8] in his extension of the result of Gács and Körner on common information [9]. Maximal correlation  $\rho(X, Y)$  of a bipartite probability distribution  $p_{XY}$  is the maximum of Pearson's correlation coefficient over all non-constant functions  $f$  and  $g$  of  $X$  and  $Y$  respectively. That is,

$$\rho(X, Y) = \max \frac{\mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])]}{\sqrt{\text{Var}[f(X)]\text{Var}[g(Y)]}}, \quad (2)$$

where  $\mathbb{E}[\cdot]$  and  $\text{Var}[\cdot]$  are expected value and variance respectively; moreover, the maximum is taken over all non-constant functions  $f = f(X)$  and  $g = g(Y)$ . Maximal correlation satisfies the following two important properties:

(i) *Tensorization*: We have

$$\rho(X_1 X_2, Y_1 Y_2) = \max\{\rho(X_1, Y_1), \rho(X_2, Y_2)\}, \quad (3)$$

when  $X_1 Y_1$  and  $X_2 Y_2$  are independent, i.e.,  $p_{X_1 X_2 Y_1 Y_2} = p_{X_1 Y_1} \cdot p_{X_2 Y_2}$ . This equation in particular gives (1).

(ii) *Monotonicity*: We have

$$\rho(A, B) \leq \rho(X, Y), \quad (4)$$

when  $A - X - Y - B$  forms a Markov chain. Thus, maximal correlation can be thought of as a *measure of correlation*

**Maximal correlation ribbon:** Another measure of correlation that satisfies the tensorization property is the *maximal correlation ribbon* (MC ribbon) defined in [3]. MC ribbon  $\mathfrak{S}(X, Y)$  is the set of  $(\lambda_1, \lambda_2) \in [0, 1]^2$  such that

$$\text{Var}[f] \geq \lambda_1 \text{Var}_X[\mathbb{E}[f|X]] + \lambda_2 \text{Var}_Y[\mathbb{E}[f|Y]], \quad (5)$$

for all functions  $f = f(X, Y)$  of both  $X$  and  $Y$ . It is shown in [3] that the MC ribbon satisfies the following properties:

(i) *Tensorization*:  $\mathfrak{S}(X_1 X_2, Y_1 Y_2) = \mathfrak{S}(X_1, Y_1) \cap \mathfrak{S}(X_2, Y_2)$  when  $X_1 Y_1$  and  $X_2 Y_2$  are independent.

(ii) *Monotonicity*:  $\mathfrak{S}(X, Y) \subseteq \mathfrak{S}(A, B)$  when  $A - X - Y - B$  forms a Markov chain.

Thus the MC ribbon satisfies properties similar to those of maximal correlation. Indeed it is shown in [3] that the maximal correlation can be characterized in terms of the MC ribbon:

$$\rho^2(X, Y) = \inf \frac{1 - \lambda_1}{\lambda_2}, \quad (6)$$

over all  $(\lambda_1, \lambda_2) \in \mathfrak{S}(X, Y)$  with  $\lambda_2 \neq 0$ . Thus the MC ribbon is a parent invariant of bipartite correlations which also characterizes  $\rho(X, Y)$ . Moreover, as will be done later in this paper, the definition (5) can easily be generalized to the multivariate case (more than two random variables).

**$\Phi$ -entropy:** Variance of a function is equal to its  $\Phi$ -entropy when we take  $\Phi(t) = t^2$ . To explain this, note that for a function  $\Phi$ , the  $\Phi$ -entropy of  $f = f(X)$  is defined by

$$H_\Phi(f) := \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}f).$$

Then for  $\Phi(t) = t^2$  we have  $H_\Phi(f) = \text{Var}[f]$ . Moreover, the MC ribbon is equal to the set of  $(\lambda_1, \lambda_1) \in [0, 1]^2$  such that for all functions  $f = f(X, Y)$  we have

$$H_\Phi(f) \geq \lambda_1 H_\Phi(\mathbb{E}[f|X]) + \lambda_2 H_\Phi(\mathbb{E}[f|Y]).$$

This expression for MC ribbon suggests generalizing it for arbitrary choices of  $\Phi$ , or at least for convex ones. This idea would seem more reasonable once we note that another important measure of correlation that satisfies the tensorization properties, namely the *hypercontractivity ribbon* (HC ribbon), can also be expressed in the above form for the choice of  $\Phi(t) = 1 - h((1+t)/2)$  where  $h(\cdot)$  is the binary entropy function:  $h(p) = -p \log p - (1-p) \log(1-p)$  (this fact is explained in details later in Example 3). As a result, the two most well-known measures of correlation that satisfy the tensorization property can be expressed in terms of  $\Phi$ -entropy as above.

**Our contributions:** Following the above ideas, for any convex function  $\Phi$  we define a  $\Phi$ -ribbon associated to any  $k$  (correlated) random variables  $(X_1, \dots, X_k)$ . We prove that  $\Phi$ -ribbon satisfies the tensorization property as well as the monotonicity property similar to the MC ribbon assuming that  $\Phi$  satisfies an important technical condition. Then the MC ribbon and the HC ribbon belong to a family of measures of correlations all of which satisfy tensorization.

The technical condition that we require  $\Phi$  to satisfy is exactly the same condition under which  $\Phi$ -entropy becomes *subadditive*. Subadditivity of entropy for *independent* random variable is a tool that is used to prove certain concentration of measure inequalities. Our  $\Phi$ -ribbon, defined for arbitrary *correlated* random variables, can be understood as a generalization of the subadditivity inequality of  $\Phi$ -entropy.

Studying  $\Phi$ -ribbon further, we show that a quantity introduced in [19], called the *strong data processing inequality constant*, can be characterized in terms of  $\Phi$ -ribbon in the same way that the MC ribbon characterizes  $\rho$ . Moreover, we show that the MC ribbon, as a set, includes all other  $\Phi$ -ribbons and in this sense is a special one. Moreover, we prove equivalent characterizations for the MC ribbon which help us to compute it more easily. In particular, we compute the MC ribbon of a multivariate Gaussian distribution in terms of its covariance matrix. We also fully characterize the MC ribbon in the bipartite case in terms of maximal correlation.

## 2 Preliminaries

Let us first fix some notations. Sets are denoted by calligraphic letters as  $\mathcal{X}$ . Random variables are denoted by capital letters as  $X$  and their values by lowercase letters as  $x \in \mathcal{X}$ . Such a random variable is determined by its distribution  $p_X$ , i.e., with values  $p(X = x) = p(x)$  for  $x \in \mathcal{X}$ . Except otherwise stated, we restrict to random variables taking values in finite sets.

We let  $[k] = \{1, 2, \dots, k\}$ . The tuple  $(\lambda_1, \lambda_2, \dots, \lambda_k)$  is sometimes denoted by  $\lambda_{[k]}$ . Similarly, when we have  $k$  random variables  $X_1, \dots, X_k$ , we use  $X_{[k]}$  to denote the tuple  $(X_1, X_2, \dots, X_k)$  for  $k \geq 1$ . When  $k = 0$ , we use  $X_{[k]}$  to denote the empty sequence. We also use  $\hat{i}$  to denote  $\{1, \dots, i-1, i+1, \dots, k\}$ , so  $X_{\hat{i}} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$ .

Let  $X$  be a random variable taking values in the finite set  $\mathcal{X}$ . Then a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  can itself be thought of as a random variable. To emphasis that  $f$  is a function of  $X$  we sometimes denoted it by  $f_X$  or  $f(X)$ . The expectation and variance of  $f$  are denoted by  $\mathbb{E}[f]$  and  $\text{Var}[f]$  respectively. We sometimes denoted them by  $\mathbb{E}_X[f]$  and  $\text{Var}_X[f]$  to emphasis that they are computed with respect to the random choice of  $X$ .

Let  $f = f_{XY} = f(X, Y)$  be a function of two random variables  $(X, Y)$  with the joint distribution  $p_{XY}$ . Then  $\mathbb{E}[f|X]$  is a function of  $X$  which is equal to the conditional expectation of  $f$ , over the random choice of  $Y$ , given a fixed value for  $X$ :

$$\mathbb{E}[f|X](x) = \mathbb{E}[f|X = x] = \sum_y p(y|x) f(x, y).$$

We can then speak of  $\text{Var}[\mathbb{E}[f|X]] = \text{Var}_X[\mathbb{E}[f|X]]$ .

A function  $\Phi$  is said to be smooth if it has derivatives of all orders everywhere in its domain.

We denote the binary entropy function by  $h(\cdot)$ , i.e.,  $h(p) = -p \log p - (1-p) \log(1-p)$  for  $p \in [0, 1]$ .

### 2.1 Hypercontractivity ribbon

We have already defined an important measure of correlation with the tensorization property in (2). Another important such measure is the *hypercontractivity ribbon* first defined by Ahlswede and Gács [17].

**Definition 1** ([17]). *The hypercontractivity ribbon (HC ribbon),  $\mathfrak{R}(X, Y)$ , associated to a pair of random variables  $(X, Y)$  is the set of all  $(\lambda_1, \lambda_2) \in [0, 1]^2$  such that for every pair of functions  $f_X$  and  $g_Y$  we have*

$$\mathbb{E}[f_X g_Y] \leq \|f_X\|_{\frac{1}{\lambda_1}} \|g_Y\|_{\frac{1}{\lambda_2}}, \quad (7)$$

where the norms  $\|\cdot\|_r$  are defined by  $\|f\|_r = \mathbb{E}[|f|^r]^{1/r}$ .

We should mention here that the HC ribbon defined in [17] is indeed the set of  $(r, s) = (1/\lambda_1, 1/\lambda_2)$  for which  $(\lambda_1, \lambda_2) \in \mathfrak{R}(X, Y)$  as we defined above. Nevertheless, we prefer this definition for later use.

HC ribbon satisfies several interesting properties for which we refer to [17]. Here we only mention the surprising result of Nair [16] that HC ribbon can be characterized in terms of mutual information (a related characterization was also found in [18]).

**Theorem 2** ([16]).  *$\mathfrak{R}(X, Y)$  is equal to the set of all pairs  $(\lambda_1, \lambda_2) \in [0, 1]^2$  such that for all  $p(u|x, y)$  we have*

$$I(XY; U) \geq \lambda_1 I(X; U) + \lambda_2 I(Y; U). \quad (8)$$

Furthermore, without loss of generality one may restrict to auxiliary random variables  $U$  that are binary.

An important quantity related to HC ribbon is the *strong data processing inequality* constant. We refer to [17] for its original definition. Here, based on the result of [20], we may define this constant  $s^*(X, Y)$ , as the smallest  $\lambda \geq 0$  such that for any  $p(u|x)$  we have

$$\lambda I(U; X) \geq I(U; Y).$$

Note that for a Markov chain  $U - X - Y$ , by the data processing inequality we have  $I(U; X) \geq I(U; Y)$  (and then  $s^*(X, Y) \leq 1$ ). That is the reason that  $s^*(X, Y)$  is called the *strong* data processing inequality constant.

$s^*(X, Y)$  can be characterized in terms of HC ribbon as follows:

$$s^*(X, Y) = \inf \frac{1 - \lambda_1}{\lambda_2},$$

where the infimum is taken over all  $(\lambda_1, \lambda_2) \in \mathfrak{R}(X, Y)$ . Observe that this characterization of  $s^*(X, Y)$  is similar to that of  $\rho(X, Y)$  given in (6).

It is straightforward to generalize the definition of HC ribbon as well as Theorem 2 to the multivariate case. The HC ribbon,  $\mathfrak{R}(X_1, \dots, X_k)$ , associated to  $k$  random variables  $X_{[k]} = (X_1, \dots, X_k)$ , is the set of tuples  $(\lambda_1, \dots, \lambda_k) \in [0, 1]^k$  such that for all functions  $f_i = f_i(X_i)$ ,  $i = 1, \dots, k$ , we have

$$\mathbb{E}[f_1 \cdots f_k] \leq \|f_1\|_{\frac{1}{\lambda_1}} \cdots \|f_k\|_{\frac{1}{\lambda_k}}.$$

Then it is easily verified that Theorem 2, with the same proof, holds in the multivariate case. That is,  $\mathfrak{R}(X_1, \dots, X_k)$  is equal to the set of tuples  $(\lambda_1, \dots, \lambda_k) \in [0, 1]^k$  such that for any auxiliary (binary) random variable  $U$  we have

$$I(X_{[k]}; U) \geq \lambda_1 I(X_1; U) + \cdots + \lambda_k I(X_k; U). \quad (9)$$

## 2.2 $\Phi$ -entropy

To present our main results we need to define and review the properties of  $\Phi$ -entropy. The reader may refer to [22, Chapter 14] for a more detailed treatment of the subject (see also [23, 24]).

Let  $f = f_X$  be a function of a random variable  $X$ . Also let  $\Phi$  be a function that is defined on a convex set that contains the range of  $f$ . Then the  $\Phi$ -entropy of  $f$  is defined by

$$H_\Phi(f) = \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}f).$$

In this paper we always assume that  $\Phi$  is convex, in which case

$$H_\Phi(f) \geq 0,$$

by Jensen's inequality. For the choice of  $\Phi(t) = t^2$ , the  $\Phi$ -entropy simply reduces to variance:  $H_\Phi(f) = \text{Var}(f)$ .

**Example 3.** Let  $p_{UA}$  be some arbitrary distribution with  $U$  taking values in  $\{+1, -1\}$ . Define  $f_A = \mathbb{E}[U|A]$ . Then  $\mathbb{E}[f] = \mathbb{E}[U]$  and we have

$$\begin{aligned} I(U; A) &= H(U) - H(U|A) \\ &= h\left(\frac{1 + \mathbb{E}f}{2}\right) - \mathbb{E}\left[h\left(\frac{1 + f}{2}\right)\right] \\ &= H_\Phi(f), \end{aligned}$$

for  $\Phi(x) = 1 - h(\frac{1+x}{2})$ , where  $h(\cdot)$  denotes the binary entropy function.

Similar to the conditional Shannon entropy, we define the conditional  $\Phi$ -entropy. Let  $f_{XY}$  be a function of two random variables  $(X, Y)$ . Then we define

$$H_\Phi(f|Y) = \mathbb{E}[\Phi(f)] - \mathbb{E}_Y[\Phi(\mathbb{E}[f|Y])] \quad (10)$$

$$= \sum_y p(y) (\mathbb{E}[\Phi(f)|Y = y] - \Phi(\mathbb{E}[f|Y = y])). \quad (11)$$

Furthermore, we set  $H_\Phi(f|Y = y) = \mathbb{E}[\Phi(f)|Y = y] - \Phi(\mathbb{E}[f|Y = y])$ , so that we have

$$H_\Phi(f|Y) = \sum_y p(y) H_\Phi(f|Y = y).$$

With these notations, we can now express  $\Phi$ -entropy's version of the *law of total variance*:

$$\begin{aligned} H_\Phi(f) &= \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}f) \\ &= \mathbb{E}[\Phi(f)] - \mathbb{E}_Y[\Phi(\mathbb{E}[f|Y])] + \mathbb{E}_Y[\Phi(\mathbb{E}[f|Y])] - \Phi(\mathbb{E}f) \\ &= H_\Phi(f|Y) + H_\Phi(\mathbb{E}[f|Y]). \end{aligned} \quad (12)$$

We call the above equation *the chain rule* for  $\Phi$ -entropy as it parallels the chain rule for Shannon entropy.

Along the same lines, one can prove the following conditional form of the chain rule for  $\Phi$ -entropy. Suppose that  $f_{XYZ}$  is a function of three random variables  $(X, Y, Z)$ . Then we have

$$H_\Phi(f|X) = H_\Phi(f|XY) + H_\Phi(\mathbb{E}[f|XY]|X), \quad (13)$$

which is a generalization of (12). This equation and the non-negativity of  $\Phi$ -entropy imply that

$$H_\Phi(f|X) \geq H_\Phi(f|XY). \quad (14)$$

In other words, just like Shannon's entropy, conditioning reduces  $\Phi$ -entropy. Observe that from the chain rule, (14) can be also written as

$$H_\Phi(\mathbb{E}[f|XY]) \geq H_\Phi(\mathbb{E}[f|X]). \quad (15)$$

Despite the above similarities between Shannon's entropy and the  $\Phi$ -entropy, one can relate  $\Phi$ -entropy to the generalized *relative entropy* of Ali-Silvey [10] and Csiszar [11][12] (also called the " $f$ -divergence"): take a non-negative function  $f(x)$  that is normalized, *i.e.*,  $\mathbb{E}_X[f] = 1$ . Then  $f(x)$  is of the form  $f(x) = q(x)/p(x)$  where  $p(x)$  is the underlying distribution on  $X$  and  $q(x)$  is some arbitrary distribution. Now,

$$H_\Phi(X) = \sum_x p(x)\Phi(f(x)) - \Phi\left(\sum_x p(x)f(x)\right) \quad (16)$$

$$= \sum_x p(x)\Phi\left(\frac{q(x)}{p(x)}\right) - \Phi(1) \quad (17)$$

is explicitly in terms of the  $\Phi$ -divergence. Ignoring the  $\Phi(1)$  term, the  $\Phi$ -entropy in (17) reduces to the KL divergence  $D(q\|p)$  for  $\Phi(x) = x \log(x)$ . Therefore,  $\Phi$ -entropy is really a relative entropy (when  $f$  is a non-negative and normalized) rather than an entropy. In fact, the analogy between Shannon's entropy and the  $\Phi$ -entropy has limitations: while Shannon entropy is *concave* in its underlying distribution, the following *convexity* property holds for the  $\Phi$ -entropy.

**Lemma 4.** *Let  $\Phi$  be a convex function and fix the distribution  $p_X$  and function  $f_X$ . Then the function*

$$p_{Y|X} \mapsto H_\Phi(\mathbb{E}[f|Y]),$$

*is convex.*

*Proof.* We note that  $H_\Phi(\mathbb{E}[f|Y]) = \mathbb{E}_Y[\Phi(\mathbb{E}[f|Y])] - \Phi(\mathbb{E}f)$ . So it suffices to prove the convexity of

$$p_{Y|X} \mapsto \mathbb{E}_Y[\Phi(\mathbb{E}[f|Y])] = \sum_y p(y)\Phi\left(\sum_x \frac{p(x)p(y|x)f(x)}{p(y)}\right),$$

which is immediate once we note that for any convex  $\Phi$ , the function

$$(s, t) \mapsto s\Phi\left(\frac{t}{s}\right),$$

is jointly convex for  $s > 0$ . □

The following simple lemma will be used frequently.

**Lemma 5.** *Let  $\Phi$  be a smooth convex function. Let  $c \in \mathbb{R}$  be point in the interior of the domain of  $\Phi$ , and  $f$  be an arbitrary function with  $\mathbb{E}f = 0$ . Then for  $g = c + \epsilon f$ , where  $|\epsilon|$  is small, we have*

$$H_\Phi(g) = \frac{1}{2}\Phi''(c)\text{Var}[f]\epsilon^2 + O(|\epsilon|^3).$$

*Proof.* Taking the Taylor expansion of  $\Phi$  around  $c$  we have

$$\Phi(c + \epsilon f) - \Phi(c) = \Phi'(c)(\epsilon f) + \frac{1}{2}\Phi''(c)(\epsilon f)^2 + O(|\epsilon|^3).$$

Now taking the expectation of both sides and noting that  $\mathbb{E}g = c$ , we obtain the desired result. □

So far the only condition we put on  $\Phi$  is convexity. We must however consider a more restricted class of functions.

**Definition 6.** We define  $\mathcal{F}$  to be the class of smooth convex functions  $\Phi$ , whose domain is a convex subset of  $\mathbb{R}$ , that are not affine (not of the form  $at + b$  for some constants  $a$  and  $b$ ) and satisfy one of the following equivalent conditions (see [22, Exercise 14.2]):

- (i)  $(s, t) \mapsto p\Phi(s) + (1 - p)\Phi(t) - \Phi(ps + (1 - p)t)$ , for any  $p \in [0, 1]$ , is jointly convex.
- (ii)  $(s, t) \mapsto \Phi(s) - \Phi(t) - \Phi'(t)(s - t)$  is jointly convex.
- (iii)  $(s, t) \mapsto (\Phi'(s) - \Phi'(t))(s - t)$  is jointly convex.
- (iv)  $(s, t) \mapsto \Phi''(s)t^2$  is jointly convex.
- (v)  $1/\Phi''$  is concave.
- (vi)  $\Phi'''\Phi'' \geq 2\Phi''^2$ .

Let us clarify a few points in this definition. First, we exclude affine functions  $\Phi(t) = at + b$  simply because  $H_\Phi(f)$  always vanishes if  $\Phi$  is affine. Second, from the above list of equivalent conditions, we mostly use (i) which has nothing to do with the smoothness of  $\Phi$ . We indeed assumed smoothness only because in this case we have the equivalent conditions (v) and (vi) which can easily be verified. Third, using (v)  $\Phi''(x)$  is strictly positive for any  $\Phi \in \mathcal{F}$ . That is, functions in  $\mathcal{F}$  are strictly convex.

Examples of functions in  $\mathcal{F}$  include  $\Phi(t) = t \log t$  and  $\Phi(t) = t^\alpha$  for  $\alpha \in (1, 2]$  as well as their affine transformations such as  $\Phi(t) = 1 - h(\frac{1+t}{2})$  and  $\Phi(t) = (1 - t)^\alpha + (1 + t)^\alpha$  for  $\alpha \in (1, 2]$ .

The following lemma is a key tool in our proofs in the next section.

**Lemma 7.** (a) Assume  $X$  and  $Y$  are independent random variables, and  $f_{XY}$  is arbitrary. Then, for any  $\Phi \in \mathcal{F}$ , we have

$$\mathbb{E}[\Phi(f)] - \mathbb{E}_X[\Phi(\mathbb{E}_Y[f|X])] \geq \mathbb{E}_Y[\Phi(\mathbb{E}_X[f|Y])] - \Phi(\mathbb{E}f),$$

or equivalently  $H_\Phi(f|X) \geq H_\Phi(\mathbb{E}[f|Y])$ .

(b) More generally, if  $f_{XYZ}$  is a function of three random variables satisfying the Markov chain condition  $X - Z - Y$ , we have

$$H_\Phi(f|XZ) \geq H_\Phi(\mathbb{E}[f|YZ]|Z).$$

(c) Under the same condition as in part (b) we have

$$H_\Phi(\mathbb{E}[f|Z]) + H_\Phi(f|XZ) \geq H_\Phi(\mathbb{E}[f|YZ]).$$

*Proof.* (a) Based on property (i) of Definition 6, an induction argument shows that for every distribution  $p_X$ , the mapping

$$f_X \mapsto \sum_x p(x)\Phi(f(x)) - \Phi\left(\sum_x p(x)f(x)\right),$$

is jointly convex. This means that for every distribution  $q_Y$  and  $f_{XY}$  we have

$$\begin{aligned} \sum_y q(y) \left( \sum_x p(x)\Phi(f(x, y)) - \Phi\left(\sum_x p(x)f(x, y)\right) \right) \\ \geq \sum_x p(x)\Phi\left(\sum_y q(y)f(x, y)\right) - \Phi\left(\sum_{x, y} p(x)q(y)f(x, y)\right). \end{aligned}$$



This is equivalent to  $H_\Phi(f|X) \geq H_\Phi(\mathbb{E}[f|Y])$ .

(b) This part is just the “conditional” version of (a). To prove this, write down the inequality of part (a) for the function  $g_{XY}^{(z)}(x, y) = f(x, y, z)$ , for every fixed  $Z = z$ , and then take average over  $z$ . Moreover, (c) follows from (b) once we use the chain rule  $H_\Phi(\mathbb{E}[f|YZ]|Z) = H_\Phi(\mathbb{E}[f|YZ]) - H_\Phi(\mathbb{E}[f|Z])$ .

□

Subadditivity is a desirable property of  $\Phi$ -entropy [22, Sec. 4.13]. Let us now prove the subadditivity of  $\Phi$ -entropy as a corollary of Lemma 7.

**Theorem 8** (Subadditivity of  $\Phi$ -entropy). *Let  $(X_1, \dots, X_k)$  be mutually independent random variables and  $f$  be an arbitrary function of them. Then for any  $\Phi \in \mathcal{F}$ , we have*

$$H_\Phi(f) \leq \sum_{i=1}^k H_\Phi(f|X_{\widehat{i}}),$$

where  $X_{\widehat{i}} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$ .

*Proof.* Recall that  $X_{[j]} = (X_1, \dots, X_j)$ . Using the conditional form of the chain rule (13) as well as part (b) of Lemma 7, for every  $0 \leq j \leq k-1$  we have

$$\begin{aligned} H_\Phi(f|X_{[j]}) &= H_\Phi(f|X_{\widehat{j+1}}) + H_\Phi(\mathbb{E}[f|X_{\widehat{j+1}}]|X_{[j]}) \\ &\leq H_\Phi(f|X_{\widehat{j+1}}) + H_\Phi(f|X_{[j+1]}). \end{aligned}$$

Summing up all these inequalities gives the desired result.

□

Observe that subadditivity of  $\Phi$ -entropy for the choice of  $\Phi(t) = t^2$  is nothing but the Efron-Stein inequality. Using the chain rule for  $\Phi$ -entropy (12),

$$H_\Phi(f) = H_\Phi(f|X_{\widehat{i}}) + H_\Phi(\mathbb{E}[f|X_{\widehat{i}}])$$

we can equivalently express the sub-additivity of  $\Phi$ -entropy as

$$H_\Phi(f) \geq \sum_{i=1}^k \frac{1}{k-1} H_\Phi(\mathbb{E}[f|X_{\widehat{i}}]). \quad (18)$$

From here, it is a short trip to motivate our notion of  $\Phi$ -ribbon, formally defined in the next section and studied in the rest of this paper. Let us ask for the set of possible non-negative coefficients  $\lambda_i$  for which

$$H_\Phi(f) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_{\widehat{i}}])$$

holds for all functions  $f$ , *i.e.*, we are asking for the best possible constants that one can substitute instead of  $1/(k-1)$  in (18). This question about the set of coefficients  $\lambda_i$  can be asked even when  $X_1, X_2, \dots, X_k$  are *correlated* sources. Letting  $Y_i = X_{\widehat{i}}$ , we can think of  $f$  as a function of  $(Y_1, Y_2, \dots, Y_k)$ , and ask for the set of coefficients  $\lambda_i$  such that

$$H_\Phi(f) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|Y_i])$$

for all functions  $f$  of  $(Y_1, Y_2, \dots, Y_k)$ . This is what we call the  $\Phi$ -ribbon associated to  $(Y_1, Y_2, \dots, Y_k)$ .

### 3 $\Phi$ -ribbon

In this section we present our main definition, namely the  $\Phi$ -ribbon, and prove some of its properties. In particular, we show that it generalizes both the MC and the HC ribbons, and satisfies the tensorization and monotonicity properties.

**Definition 9.** Let  $\Phi \in \mathcal{F}$ . For arbitrarily distributed random variables  $(X_1, X_2, \dots, X_k)$  we define its  $\Phi$ -ribbon, denoted by  $\mathfrak{R}_\Phi(X_1, \dots, X_k) = \mathfrak{R}(X_{[k]})$ , to be the set of all  $k$ -tuples  $(\lambda_1, \lambda_2, \dots, \lambda_k)$  of non-negative numbers such that for every function  $f_{X_{[k]}}$  we have

$$H_\Phi(f) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_i]). \quad (19)$$

Note that we require the above equation for any function  $f_{X_{[k]}}$  whose range is in the domain of  $\Phi \in \mathcal{F}$ .

From the definition it is clear that  $\Phi$ -ribbon for the choice of  $\Phi(t) = t^2$  reduces to the MC ribbon defined in the Introduction. Furthermore, according to Example 3 and Theorem 2, if  $\Phi(t) = 1 - h(\frac{1+t}{2})$ , the ribbon  $\mathfrak{R}_\Phi(X_{[k]})$  becomes the HC ribbon.

Letting  $f$  to be only a function of  $X_i$ , for some  $i \in [k]$ , we find that  $\mathbb{E}[f|X_i] = f$  and hence  $H_\Phi(\mathbb{E}[f|X_i]) = H_\Phi(f)$ . Then, for any  $(\lambda_1, \dots, \lambda_k) \in \mathfrak{R}_\Phi(X_1, \dots, X_k)$  we must have  $\lambda_i \leq 1$ ; that is,

$$\mathfrak{R}_\Phi(X_1, \dots, X_k) \subseteq [0, 1]^k.$$

On the other hand, using the chain rule and the fact that  $\Phi$ -entropy is non-negative, we have  $H_\Phi(f) \geq H_\Phi(\mathbb{E}[f|X_i])$  for every  $i$ . Therefore, the tuple  $(\lambda_1, \dots, \lambda_k)$  of non-negative numbers, always belongs to  $\mathfrak{R}_\Phi(X_{[k]})$  if  $\sum_i \lambda_i \leq 1$ . This means that the nontrivial part of the  $\Phi$ -ribbon is the set of  $k$ -tuples of non-negative numbers whose sum is greater than one.

**Example 10.** If  $X_1 = X_2 = \dots = X_k$  are non-constant, then  $\mathbb{E}[f|X_i] = f$  and  $H_\Phi(f) = H_\Phi(\mathbb{E}[f|X_i])$ , for every  $i$ . As a result,  $\mathfrak{R}_\Phi(X_{[k]})$  contains only those  $(\lambda_1, \lambda_2, \dots, \lambda_k) \in [0, 1]^k$  that satisfy  $\sum_i \lambda_i \leq 1$ .

In the following we show that the  $\Phi$ -ribbon is the largest possible ribbon when  $X_i$ 's are mutually independent.

**Proposition 11.** For any  $\Phi \in \mathcal{F}$ , if  $(X_1, \dots, X_k)$  are mutually independent, we have  $\mathfrak{R}_\Phi(X_{[k]}) = [0, 1]^k$ .

*Proof.* We need to show that

$$H_\Phi(f) \geq \sum_{i=1}^k H_\Phi(\mathbb{E}[f|X_i]). \quad (20)$$

For this, we again use the conditional form of the chain rule (13) as well as part (b) of Lemma 7. For any  $0 \leq j \leq k-1$  we have

$$\begin{aligned} H_\Phi(f|X_{[j]}) &= H_\Phi(f|X_{[j+1]}) + H_\Phi(\mathbb{E}[f|X_{[j+1]}]|X_{[j]}) \\ &\geq H_\Phi(f|X_{[j+1]}) + H_\Phi(\mathbb{E}[f|X_{j+1}]), \end{aligned}$$

where in the second line we use Lemma 7 for the function  $\mathbb{E}[f|X_{[j+1]}]$ . Summing up all these inequalities gives the desired result.  $\square$

We can now prove the main result of this section, namely the tensorization and monotonicity properties of the HC ribbon and MC ribbon extend to the  $\Phi$ -ribbon.

**Theorem 12.** *For any  $\Phi \in \mathcal{F}$ , the  $\Phi$ -ribbon satisfies monotonicity and tensorization as follows:*

- (i) *Data processing: if  $(X_{[k]}, Y_{[k]})$  are random variables whose joint probability distribution satisfies  $p(y_{[k]}|x_{[k]}) = \prod_{i=1}^n p(y_i|x_i)$ , then*

$$\mathfrak{R}_\Phi(X_{[k]}) \subseteq \mathfrak{R}_\Phi(Y_{[k]}).$$

- (ii) *Tensorization: if  $X_{[k]}$  are independent of  $Y_{[k]}$ , i.e.,  $p(x_{[k]}, y_{[k]}) = p(x_{[k]})p(y_{[k]})$ , then*

$$\mathfrak{R}_\Phi(X_1 Y_1, X_2 Y_2, \dots, X_k Y_k) = \mathfrak{R}_\Phi(X_1, X_2, \dots, X_k) \cap \mathfrak{R}_\Phi(Y_1, Y_2, \dots, Y_k).$$

*Proof.* (i) Let  $\lambda_{[k]} \in \mathfrak{R}_\Phi(X_{[k]})$ , we show that  $\lambda_{[k]} \in \mathfrak{R}_\Phi(Y_{[k]})$ . For this we need to show that for every function  $f_{Y_{[k]}}$  of  $Y_{[k]}$  we have

$$H_\Phi(f) \geq \sum_i \lambda_i H_\Phi(\mathbb{E}[f|Y_i]).$$

Using the definition of  $\lambda_{[k]} \in \mathfrak{R}_\Phi(X_{[k]})$  applied to the function  $\mathbb{E}[f|X_{[k]}]$  we find that

$$H_\Phi(\mathbb{E}[f|X_{[k]}]) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_i]). \quad (21)$$

On the other hand, since  $Y_{[k]}$  are mutually independent conditioned on  $X_{[k]}$ , using Proposition 11 (in fact the “conditional” version of (20)) we find that

$$H_\Phi(f|X_{[k]}) \geq \sum_{i=1}^k H_\Phi(\mathbb{E}[f|X_{[k]}Y_i]|X_{[k]}) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_{[k]}Y_i]|X_{[k]}). \quad (22)$$

Summing up (21) and (22), and using the chain rule we arrive at

$$H_\Phi(f) \geq \sum_{i=1}^k \lambda_i \left( H_\Phi(\mathbb{E}[f|X_i]) + H_\Phi(\mathbb{E}[f|X_{[k]}Y_i]|X_{[k]}) \right).$$

Therefore, it suffices to verify that  $H_\Phi(\mathbb{E}[f|X_i]) + H_\Phi(\mathbb{E}[f|X_{[k]}Y_i]|X_{[k]}) \geq H_\Phi(\mathbb{E}[f|Y_i])$  for every  $i$ . For a fixed  $i$ , let  $g_{X_{[k]}Y_i} = \mathbb{E}[f|X_{[k]}Y_i]$ . Then this inequality can be written as

$$H_\Phi(\mathbb{E}[g|X_i]) + H_\Phi(g|X_{[k]}) \geq H_\Phi(\mathbb{E}[g|Y_i]). \quad (23)$$

Now note that we have the Markov chain  $X_{\hat{i}} - X_i - Y_i$ , so by part (c) of Lemma 7 we have

$$H_\Phi(\mathbb{E}[g|X_i]) + H_\Phi(g|X_{[k]}) \geq H_\Phi(\mathbb{E}[g|X_iY_i]).$$

Then (23) follows from (15).

- (ii) In the definition of  $\mathfrak{R}_\Phi(X_1 Y_1, X_2 Y_2, \dots, X_k Y_k)$  by restricting to functions  $f_{X_{[k]}}$  of  $X_{[k]}$  only, or to functions  $f_{Y_{[k]}}$  of  $Y_{[k]}$  only, we find that

$$\mathfrak{R}_\Phi(X_1 Y_1, X_2 Y_2, \dots, X_k Y_k) \subseteq \mathfrak{R}_\Phi(X_1, X_2, \dots, X_k) \cap \mathfrak{R}_\Phi(Y_1, Y_2, \dots, Y_k).$$

To prove the inclusion in the other direction, let

$$(\lambda_1, \dots, \lambda_k) \in \mathfrak{R}_\Phi(X_1, X_2, \dots, X_k) \cap \mathfrak{R}_\Phi(Y_1, Y_2, \dots, Y_k),$$

and let  $f_{X_{[k]}Y_{[k]}}$  be arbitrary. We need to show that

$$H_\Phi(f) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_i Y_i]). \quad (24)$$

Using our assumption on  $(\lambda_1, \dots, \lambda_k)$  for function  $\mathbb{E}[f|X_{[k]}]$ , as a function of  $X_{[k]}$ , we have

$$H_\Phi(\mathbb{E}[f|X_{[k]}]) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_i]).$$

Next considering  $f$ , for a fixed  $X_{[k]} = x_{[k]}$ , as a function of  $Y_{[k]}$ . Observe that conditioned on  $X_{[k]} = x_{[k]}$  the distribution of  $Y_{[k]}$  does not change. Therefore, since  $(\lambda_1, \dots, \lambda_k)$  belongs to  $\mathfrak{R}(Y_1, \dots, Y_k)$  we have

$$H_\Phi(f|X_{[k]}) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_{[k]} Y_i]|X_{[k]}).$$

Summing up these two inequalities and using chain rule, (24) is implied if we have

$$H_\Phi(\mathbb{E}[f|X_i]) + H_\Phi(\mathbb{E}[f|X_{[k]} Y_i]|X_{[k]}) \geq H_\Phi(\mathbb{E}[f|X_i Y_i]), \quad \forall i.$$

This inequality follows once we note that we have the Markov chain  $X_{\hat{i}} - X_i - Y_i$ , and we can use part (c) of Lemma 7 for the function  $\mathbb{E}[f|X_{[k]} Y_i]$ . □

Theorem 2 provides a description of the HC ribbon in terms of mutual information. Given  $p(x, y)$ , one can define the  $\Phi$ -mutual information between  $X$  and  $Y$  as follows [19]:

$$I_\Phi(X; Y) = \sum_{x, y} p(x)p(y) \Phi\left(\frac{p(x, y)}{p(x)p(y)}\right) - \Phi(1).$$

This definition differs from the one used in [19] due to the subtraction of the  $\Phi(1)$  term. This subtraction ensures that  $I_\Phi(X; Y) = 0$  when  $X$  and  $Y$  are independent. Observe that  $\Phi$ -mutual information reduces to Shannon's mutual information for  $\Phi(x) = x \log(x) \in \mathcal{F}$ . Then, similar to the statement of Theorem 2, we may define the  $I_\Phi$ -ribbon as follows:

**Definition 13.** Let  $\Phi \in \mathcal{F}$ . For arbitrarily distributed random variables  $(X_1, X_2, \dots, X_k)$  we define its  $I_\Phi$ -ribbon, denoted by  $\mathfrak{R}_{I_\Phi}(X_{[k]})$ , to be the set of all  $k$ -tuples  $(\lambda_1, \lambda_2, \dots, \lambda_k)$  of non-negative numbers such that for any  $p(u|x_{[k]})$ , we have

$$\sum_i \lambda_i I_\Phi(U; X_i) \leq I_\Phi(U; X_{[k]}).$$

Then, we have the following:

**Theorem 14.** We have  $\mathfrak{R}_{I_\Phi}(X_{[k]}) = \mathfrak{R}'_\Phi(X_{[k]})$  where  $\mathfrak{R}'_\Phi$  is the set of  $(\lambda_1, \dots, \lambda_k)$  such that for any  $f(x_{[k]})$  satisfying  $f \geq 0$  and  $\mathbb{E}[f] = 1$ , we have

$$\sum_i \lambda_i H_\Phi(\mathbb{E}[f|X_i]) \leq H_\Phi(f).$$

Observe that  $\mathfrak{R}'_\Phi$  has the same definition as  $\mathfrak{R}_\Phi$ , except that in  $\mathfrak{R}'_\Phi$  we restrict to functions  $f(x_{[k]})$  satisfying  $f \geq 0$  and  $\mathbb{E}[f] = 1$ .

The condition  $f \geq 0$  and  $\mathbb{E}[f] = 1$  essentially says that  $f(x_{[k]})$  can be written as  $q(x_{[k]})/p(x_{[k]})$  where  $p(x_{[k]})$  is the given distribution on  $X_{[k]}$  and  $q(x_{[k]})$  is some arbitrary distribution. The proof of this theorem is given in Appendix B.

### 3.1 Examples

Natural choices for the function  $\Phi(t) \in \mathcal{F}$  are

$$\varphi_\alpha(t) = t^\alpha, \quad \alpha \in (1, 2].$$

Note that  $\varphi(x)$  is defined only for  $t \geq 0$ . Without changing the corresponding  $\Phi$ -ribbon, we can even restrict the domain of  $\varphi_\alpha$  to  $[0, 1]$  simply because  $\varphi_\alpha(ct) = c^\alpha \varphi_\alpha(t)$  implies that the equation  $H_\Phi(f) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_i])$  holds for  $f$  if and only if it holds for a scaled version of  $f$ .

Another special choice of interest is  $\Phi_1(t) = 1 - h(\frac{1+t}{2})$  which, as mentioned before, results in the HC ribbon. One way to understand this function is to consider the class of functions  $\Phi_\alpha$  for  $\alpha \in (1, 2]$  defined by

$$\Phi_\alpha(t) = \frac{(1+t)^\alpha + (1-t)^\alpha - 2}{2^\alpha - 2}, \quad \alpha \in (1, 2], \quad t \in [-1, 1]. \quad (25)$$

Then  $\Phi_\alpha \in \mathcal{F}$ , and we have

$$\lim_{\alpha \searrow 1} \Phi_\alpha(t) = \Phi_1(t), \quad \forall t \in [-1, 1].$$

Note that  $\Phi_2 = \varphi_2$  for which the associated  $\Phi$ -ribbon is equal to the MC ribbon.

**Theorem 15.** *For all  $\alpha \in (1, 2]$  we have  $\mathfrak{R}_{\varphi_\alpha}(X_1, \dots, X_k) = \mathfrak{R}_{\Phi_\alpha}(X_1, \dots, X_k)$ .*

A consequence of the above theorem is that by varying the parameter  $\alpha$  from 1 to 2, the  $\Phi$ -ribbon varies from the HC ribbon to the MC ribbon. The proof of this theorem is given in Appendix C.

## 4 Strong data processing inequalities

Let us focus on the bipartite case, namely,  $k = 2$ . Suppose that we have two random variables  $(X, Y)$ . Then for any function  $f_X$  of  $X$ , we have

$$H_\Phi(f) \geq H_\Phi(\mathbb{E}[f|Y]).$$

This inequality can be thought of as a data processing inequality. Indeed, when  $\Phi(t) = 1 - h((1+t)/2)$ , according to Example 3, this inequality is equivalent to  $I(U; X) \geq I(U; Y)$  assuming the Markov chain  $U - X - Y$ . Here, we are interested in how tight this inequality is.

**Definition 16.** *For any convex function  $\Phi \in \mathcal{F}$  define  $\eta_\Phi(X, Y)$  to be the smallest (the infimum of)  $\lambda \geq 0$  such that for any function  $f_X$  (whose range is in the domain of  $\Phi$ ) we have*

$$\lambda H_\Phi(f) \geq H_\Phi(\mathbb{E}[f|Y]).$$

*We call  $\eta_\Phi(X, Y)$  the  $\Phi$ -strong data processing inequality constant ( $\Phi$ -SDPI constant).*

We borrowed the term  $\Phi$ -SDPI constant from Raginsky [19] who defines almost the same invariant. The only difference is that in the definition of [19] it is assumed that  $\mathbb{E}[f] = 1$ . This extra assumption, however, does not make a difference in the interesting example of  $\Phi(t) = t^\alpha$  as  $f_X$  can be scaled (as mentioned in Section 3.1).

From the convexity of  $\Phi$ , it is clear that  $\eta_\Phi(X, Y) \in [0, 1]$ . Moreover, if  $X$  and  $Y$  are independent we have  $\eta_\Phi(X, Y) = 0$ . Also,  $\eta_\Phi(X, Y) = 1$  if  $X = Y$ , or more generally if  $X$  and  $Y$  have explicit common data ( $f(X) = g(Y)$  for some non-constant  $f$  and  $g$ ).

When  $\Phi(t) = t^2$ ,  $\eta_\Phi(X, Y)$  is nothing but  $\rho^2(X, Y)$  as shown in [6, 7]. Moreover, for the choice of  $\Phi(t) = 1 - h((1+t)/2)$ ,  $\eta_\Phi(X, Y)$  is known [20] to be equal to  $s^*(X, Y)$  defined in [17].

**Example 17.** *The doubly symmetric binary source with parameter  $\lambda$  denoted by  $DSPB(\lambda)$  is defined as follows:  $X$  and  $Y$  are binary and uniform, and*

$$p(X = 0, Y = 1) = p(X = 1, Y = 0) = \frac{1 - \lambda}{4}.$$

*It is known that  $\rho^2(X, Y) = s^*(X, Y) = \lambda^2$ . We will show in Appendix A that*

$$\eta_\Phi(X, Y) = \lambda^2,$$

*holds for all  $\Phi \in \mathcal{F}$ .*

Let us start investigating properties of  $\eta_\Phi(X, Y)$  with two results already proved in [19]. The proof of the following proposition is an immediate consequence of Lemma 4.

**Proposition 18** ([19]). *For any convex  $\Phi$ , and fix  $p_X$  the function*

$$p_{Y|X} \mapsto \eta_\Phi(X, Y),$$

*is convex.*

It is well-known that  $s^*(X, Y) \geq \rho^2(X, Y)$ . The following theorem is a generalization of this fact.

**Theorem 19** ([19]). *For any  $\Phi \in \mathcal{F}$  we have*

$$\eta_\Phi(X, Y) \geq \rho^2(X, Y),$$

*where  $\rho(X, Y)$  is the maximal correlation.*

*Proof.* Since  $\rho^2(X, Y) = \eta_\Psi(X, Y)$  for  $\Psi(t) = t^2$  we need to show that for any  $f_X$  with  $\mathbb{E}[f] = 0$  we have

$$\eta_\Phi(X, Y) \text{Var}[f] \geq \text{Var}[\mathbb{E}[f|Y]]. \quad (26)$$

Take some  $c$  in the interior of the domain of  $\Phi$ , and consider the function  $g_X = c + \epsilon f_X$  for small  $|\epsilon| > 0$ . Then by definition we have

$$\eta_\Phi(X, Y) H_\Phi(g) \geq H_\Phi(\mathbb{E}[g|Y]).$$

Now 26 follows by applying Lemma 5 on both sides of this inequality. □

We can now state our main result about the  $\Phi$ -SDPI constant.

**Theorem 20.** Let  $\Phi \in \mathcal{F}$  be a convex function defined on some compact interval. Then we have

$$\eta_\Phi(X, Y) = \inf \frac{1 - \lambda_1}{\lambda_2}, \quad (27)$$

where the infimum is taken over all  $(\lambda_1, \lambda_2) \in \mathfrak{R}_\Phi(X, Y)$  with  $\lambda_2 \neq 0$ .

This theorem for  $\Phi(t) = 1 - h((1+t)/2)$  is derived from the results of [17] and [20]. For  $\Phi(t) = t^2$  it is proved in [3].

This theorem has a technical assumption, that the domain of  $\Phi$  is compact. We do not know whether the theorem holds without this restriction. Nevertheless, this assumption is already satisfied (or can be assumed without loss of generality) for all examples given in Section 3.1. For instance, for  $\Phi(t) = t^2$ , without loss of generality we can restrict the domain of  $\Phi$  to  $[-1, 1]$  by properly scaling the function  $f$ . To see this observe that the equation  $H_\Phi(f) \geq \sum_{i=1}^k \lambda_i H_\Phi(\mathbb{E}[f|X_i])$  holds for  $f$  if and only if it holds for a scaled version of  $f$ .

*Proof.* Let  $(\lambda_1, \lambda_2) \in \mathfrak{R}_\Phi(X, Y)$  with  $\lambda_2 \neq 0$ . Then for any function  $f_X$  of  $X$  we have

$$H_\Phi(f) \geq \lambda_1 H_\Phi(\mathbb{E}[f|X]) + \lambda_2 H_\Phi(\mathbb{E}[f|Y]).$$

Since  $f$  is taken to be a function of  $X$  only, we have  $f = \mathbb{E}[f|X]$ . Therefore,

$$\frac{1 - \lambda_1}{\lambda_2} H_\Phi(f) \geq H_\Phi(\mathbb{E}[f|Y]).$$

Thus we have  $(1 - \lambda_1)/\lambda_2 \geq \eta_\Phi(A, B)$ , and then

$$\eta_\Phi(X, Y) \leq \inf \frac{1 - \lambda_1}{\lambda_2}.$$

To prove the inequality in the other direction we show that for any  $\delta > 0$ , we have

$$\eta_\Phi(A, B) + \delta \geq \inf \frac{1 - \lambda_1}{\lambda_2}.$$

To show this, it suffices to argue that there exists  $n$  such that the pair  $(\lambda_1^{(n)}, \lambda_2^{(n)})$  given by

$$\lambda_1^{(n)} = 1 - \frac{\eta_\Phi + \delta}{n}, \quad \lambda_2^{(n)} = \frac{1}{n},$$

where  $\eta_\Phi = \eta_\Phi(X, Y)$ , belongs to  $\mathfrak{R}_\Phi(X, Y)$ . Suppose that this is not the case. Then, for any  $n$  there is a function  $f_{AB}^{(n)}$  such that

$$H_\Phi(f^{(n)}) < \lambda_1^{(n)} H_\Phi(\mathbb{E}[f^{(n)}|X]) + \lambda_2^{(n)} H_\Phi(\mathbb{E}[f^{(n)}|Y]).$$

Using the chain rule, this inequality can be rewritten as

$$\frac{1 - \lambda_1^{(n)}}{\lambda_2^{(n)}} H_\Phi(\mathbb{E}[f^{(n)}|X]) + \frac{1}{\lambda_2^{(n)}} H_\Phi(f^{(n)}|X) < H_\Phi(\mathbb{E}[f^{(n)}|Y]),$$

or equivalently as

$$(\eta_\Phi + \delta) H_\Phi(\mathbb{E}[f^{(n)}|X]) + n H_\Phi(f^{(n)}|X) < H_\Phi(\mathbb{E}[f^{(n)}|Y]).$$

Since  $\Phi$ -entropy is non-negative, we infer from this inequality that

$$(\eta_\Phi + \delta)H_\Phi(\mathbb{E}[f^{(n)}|X]) < H_\Phi(\mathbb{E}[f^{(n)}|Y]), \quad (28)$$

and

$$nH_\Phi(f^{(n)}|X) < H_\Phi(\mathbb{E}[f^{(n)}|Y]). \quad (29)$$

Since  $\mathcal{X}$  and  $\mathcal{Y}$  are assumed to be finite, and the images of functions  $f^{(n)}$  are in a compact interval (i.e., the domain of  $\Phi$ ), there is an increasing sequence  $\{n_k : k \geq 1\}$  such that

$$\lim_{k \rightarrow \infty} f^{(n_k)}(x, y) = \hat{f}(x, y), \quad \forall x, y,$$

for some function  $\hat{f}$ .

$\Phi$  is continuous and defined on a compact interval. Thus, there is a constant  $M > 0$ , such that  $|\Phi(t)| \leq M$  for all  $t$ . As a result, the  $\Phi$ -entropy of any function is at most  $2M$ . Then from (29) we have

$$H_\Phi(f^{(n)}|X) \leq \frac{2M}{n}, \quad \forall n. \quad (30)$$

Then, by a continuity argument we conclude that  $H_\Phi(\hat{f}|X) = 0$ , i.e.,  $\mathbb{E}[\Phi(\hat{f})] = \mathbb{E}_X[\Phi(\mathbb{E}[\hat{f}|X])]$ . From this equality and the fact that  $\Phi \in \mathcal{F}$  is strictly convex, we infer that  $\hat{f} = \hat{f}_X$  is a function of  $X$  only.

Next, using (28) we find that

$$(\eta_\Phi + \delta)H_\Phi(\mathbb{E}[\hat{f}|X]) = (\eta_\Phi + \delta)H_\Phi(\hat{f}) \leq H_\Phi(\mathbb{E}[\hat{f}|Y]).$$

Moreover, since  $\hat{f}$  is a function of  $X$ , by the definition of  $\eta_\Phi(X, Y)$  we have

$$H_\Phi(\mathbb{E}[\hat{f}|Y]) \leq \eta_\Phi H_\Phi(\hat{f}).$$

Putting these two inequalities together we conclude that  $H_\Phi(\hat{f}) = 0$ , which again by the strict convexity of  $\Phi$  imply that  $\hat{f}$  is a constant, i.e.,

$$\lim_{k \rightarrow \infty} f^{(n_k)} = c,$$

for some constant  $c$ .

Observe that  $\mathbb{E}[f^{(n)}|X]$  is not a constant since otherwise  $H_\Phi(\mathbb{E}[f^{(n)}|X]) = 0$  and  $H_\Phi(f^{(n)}) = H_\Phi(f^{(n)}|X)$ . In this case, from (29) we find that  $nH_\Phi(\mathbb{E}[f^{(n)}|Y]) \leq nH_\Phi(f^{(n)}) < H_\Phi(\mathbb{E}[f^{(n)}|Y])$  which is a contradiction. As a result,  $\mathbb{E}[f^{(n)}|X]$  is not a constant and  $\text{Var}_X \mathbb{E}[f^{(n)}|X] > 0$ . Then for every  $k$  we can write

$$f^{(n_k)} = c_k + \epsilon_k g^{(k)},$$

such that

$$c_k = \mathbb{E}[f^{(n_k)}], \quad \mathbb{E}[g^{(k)}] = 0, \quad \epsilon_k = \sqrt{\text{Var}_X \mathbb{E}[f^{(n_k)}|X]} > 0, \quad \text{Var}_X \mathbb{E}[g^{(k)}|X] = 1.$$

Observe that  $\lim_{k \rightarrow \infty} \epsilon_k = 0$  since  $f^{(n_k)}$ , and then  $\mathbb{E}[f^{(n_k)}|X]$  converge to constant functions. We also have  $\lim_{k \rightarrow \infty} c_k = c$ . Moreover,  $g^{(k)}$ 's are uniformly bounded since they have zero expectation and  $\text{Var}_X \mathbb{E}[g^{(k)}|X] = 1$ .



Now using Lemma 5 we find that

$$\left| H_\Phi(f^{(n_k)}) - \frac{1}{2}\Phi''(c_k)\text{Var}[g^{(k)}]\epsilon_k^2 \right| = O(\epsilon_k^3).$$

Here, in particular, we use the fact that  $g^{(k)}$ 's are uniformly bounded. We similarly have

$$\left| H_\Phi(\mathbb{E}[f^{(n_k)}|X]) - \frac{1}{2}\Phi''(c_k)\text{Var}_X[\mathbb{E}[g^{(k)}|X]]\epsilon_k^2 \right| = O(\epsilon_k^3),$$

and

$$\left| H_\Phi(\mathbb{E}[f^{(n_k)}|Y]) - \frac{1}{2}\Phi''(c_k)\text{Var}_Y[\mathbb{E}[g^{(k)}|Y]]\epsilon_k^2 \right| = O(\epsilon_k^3),$$

Using these in (28) and (30), and noting that  $\epsilon_k > 0$  and that  $0 < \Phi''(c_k) < \Phi''(c) + 1$  for sufficiently large  $k$ , we find that

$$(\eta_\Phi + \delta)\text{Var}_X[\mathbb{E}[g^{(k)}|X]] < \text{Var}_Y[\mathbb{E}[g^{(k)}|Y]] + O(\epsilon_k), \quad (31)$$

and

$$\text{Var}[g^{(k)}|X] < \frac{M'}{n_k} + O(\epsilon_k), \quad (32)$$

for some constant  $M'$ .

As mentioned above the functions  $g^{(k)}$  are uniformly bounded. Then there is an increasing sequence  $\{k_j : j \geq 1\}$  such that

$$\lim_{j \rightarrow \infty} g^{(k_j)} = \hat{g},$$

for some function  $g_{XY}$ . Then using (32) we have  $\text{Var}[\hat{g}|X] = 0$ , i.e.,  $\hat{g} = \hat{g}_X$  is a function of  $X$  only. On the other hand, since  $\text{Var}[g^{(k)}|X] = 1$ , for all  $k$ , we have  $\text{Var}[\hat{g}|X] = \text{Var}[\hat{g}] = 1$ , i.e.,  $\hat{g}$  is not a constant.

Next using (31) we find that

$$(\eta_\Phi + \delta)\text{Var}[\hat{g}] \leq \text{Var}[\mathbb{E}[\hat{g}|Y]].$$

Also, since  $\rho(X, Y) = \eta_\Psi(X, Y)$  for  $\Psi(t) = t^2$ , and  $\hat{g}$  is a function of  $X$  we have

$$\text{Var}[\mathbb{E}[\hat{g}|Y]] \leq \rho^2(X, Y)\text{Var}[\hat{g}].$$

Comparing the above two inequalities and using  $\text{Var}[\hat{g}] = 1$  we conclude that  $\eta_\Phi(X, Y) + \delta \leq \rho^2(X, Y)$ , which is in contradiction with Theorem 19. We are done.  $\square$

We now state the tensorization and monotonicity properties of the  $\Phi$ -SDPI constant.

**Theorem 21.** *For any  $\Phi \in \mathcal{F}$ , the  $\Phi$ -SDPI constant  $\eta_\Phi(X, Y)$  satisfies the followings:*

- (i) *Monotonicity: If  $X - A - B - Y$  forms a Markov chain, then  $\eta_\Phi(X, Y) \leq \eta_\Phi(A, B)$ .*
- (ii) *Tensorization: If  $p_{ABXY} = p_{AB} \cdot p_{XY}$  then*

$$\eta_\Phi(AX, BY) = \max \{ \eta_\Phi(A, B), \eta_\Phi(X, Y) \}.$$

The tensorization of  $\eta_\Phi(X, Y)$  is already proved in [19]. Moreover, this theorem, in the case that the domain of  $\Phi$  is compact, is a simple corollary of Theorem 20 and the monotonicity and tensorization properties of the  $\Phi$ -ribbon.

In Appendix A we give a generalization of the SDPI constant associated to two functions  $\Phi, \Psi$ . This constant which we denote by  $\eta_{\Phi, \Psi}(X, Y)$ , coincides with  $\eta_\Phi(X, Y)$  when  $\Phi = \Psi$ . We prove in Appendix A that  $\eta_{\Phi, \Psi}(X, Y)$  satisfies the tensorization and monotonicity properties, from which Theorem 21 follows as a special case.

#### 4.1 Example: sums of i.i.d. random variables

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. random variables. For any  $1 \leq \ell \leq n$  define

$$S_\ell = X_1 + \dots + X_\ell.$$

It is known [13] that for any  $1 \leq m \leq n$  we have  $\rho^2(S_n, S_m) = \frac{m}{n}$ . Moreover, recently it is shown in [14] that  $s^*(S_n, S_m) = \frac{m}{n}$ . In the following we prove a similar result for all  $\Phi \in \mathcal{F}$ .

**Theorem 22.** *Let  $X_1, \dots, X_n$  be in i.i.d. random variables. Then for any  $1 \leq m \leq n$  and any  $\Phi \in \mathcal{F}$  we have*

$$\eta_\Phi(S_n, S_m) = \frac{m}{n}.$$

To prove this theorem we borrow ideas from [14].

*Proof.* By Theorem 19 we already know that  $\eta_\Phi(S_n, S_m) \geq \rho^2(S_n, S_m) = m/n$ . Then it suffices to show that  $\eta_\Phi(S_n, S_m) \leq m/n$ . For this we need to verify that for any function  $f = f(S_n)$  we have

$$\frac{m}{n} H_\Phi(f) \geq H_\Phi(\mathbb{E}[f|S_m]). \quad (33)$$

Observe that for any  $1 \leq \ell \leq n$ , the conditional distribution of  $S_n$  given  $X_{[\ell]} = x_{[\ell]}$ , for any  $(x_1, \dots, x_\ell)$ , is identical to the conditional distribution of  $S_n$  given  $S_\ell = x_1 + \dots + x_\ell$ . Therefore, we have  $H_\Phi(f|X_{[\ell]}) = H_\Phi(f|S_\ell)$ . Then using the chain rule

$$H_\Phi(f) = H_\Phi(\mathbb{E}[f|X_{[\ell]}]) + H_\Phi(f|X_{[\ell]}) = H_\Phi(\mathbb{E}[f|S_\ell]) + H_\Phi(f|X_\ell),$$

we find that

$$H_\Phi(\mathbb{E}[f|X_{[\ell]}]) = H_\Phi(\mathbb{E}[f|S_\ell]).$$

Let us denote the above quantity by  $c_\ell$ . We claim that

$$c_{\ell+1} - c_\ell \geq c_\ell - c_{\ell-1}, \quad 0 \leq \ell \leq n. \quad (34)$$

To prove our claim, note that since  $X_i$ 's are i.i.d. we have  $H_\Phi(\mathbb{E}[f|X_{[\ell]}]) = H_\Phi(\mathbb{E}[f|X_{[\ell-1]}, X_{\ell+1}])$ . Therefore, by the chain rule we have

$$\begin{aligned} c_{\ell+1} - 2c_\ell + c_{\ell-1} &= H_\Phi(\mathbb{E}[f|X_{[\ell-1]}, X_\ell, X_{\ell+1}]) - H_\Phi(\mathbb{E}[f|X_{[\ell-1]}, X_{\ell+1}]) \\ &\quad - H_\Phi(\mathbb{E}[f|X_{[\ell-1]}, X_\ell]) + H_\Phi(\mathbb{E}[f|X_{[\ell-1]}]) \\ &= H_\Phi(\mathbb{E}[f|X_{[\ell-1]}, X_\ell, X_{\ell+1}]|X_{[\ell-1]}, X_{\ell+1}) \\ &\quad - H_\Phi(\mathbb{E}[f|X_{[\ell-1]}, X_\ell]|X_{[\ell-1]}). \end{aligned}$$

Let us define  $g = \mathbb{E}[f|X_{[\ell-1]}, X_\ell, X_{\ell+1}]$ . Then we have

$$c_{\ell+1} - 2c_\ell + c_{\ell-1} = H_\Phi(g|X_{[\ell-1]}, X_{\ell+1}) - H_\Phi(\mathbb{E}[g|X_{[\ell-1]}, X_\ell]|X_{[\ell-1]}).$$

Now since  $X_{[\ell-1]}, X_\ell$  and  $X_{\ell+1}$  are independent, using part (b) of Lemma 7 we arrive at  $c_{\ell+1} - 2c_\ell + c_{\ell-1} \geq 0$  and then (34).

We prove by induction that

$$\frac{c_\ell}{\ell} \geq \frac{c_{\ell-1}}{\ell-1}.$$

The base case  $\ell = 2$  is immediate from (34) and that  $c_0 = 0$ . The induction step follows from

$$c_{\ell+1} - c_\ell \geq c_\ell - c_{\ell-1} \geq c_\ell - \frac{\ell-1}{\ell} c_\ell = \frac{1}{\ell} c_\ell.$$

We conclude that  $c_n/n \geq c_m/m$  since  $m \leq n$ , which is equivalent to (33). □

## 5 Maximal correlation ribbon

In this section we focus on the function  $\Phi(x) = x^2$ . We note that  $\Phi \in \mathcal{F}$ , so the ribbon  $\mathfrak{R}_\Phi(X_{[k]})$  satisfies monotonicity and tensorization. The ribbon  $\mathfrak{R}_\Phi(X_{[k]})$  for this particular function is introduced in [3] as the maximal correlation ribbon (MC ribbon) and is denoted by  $\mathfrak{S}(X_{[k]})$ . So we will use this notation here too. Thus  $\mathfrak{S}(X_{[k]})$  is the set of  $k$ -tuples  $(\lambda_1, \dots, \lambda_k) \in [0, 1]^k$  such that for all functions  $f_{X_{[k]}}$  we have

$$\text{Var}[f] \geq \sum_{i=1}^k \lambda_i \text{Var}_{X_i} [\mathbb{E}[f|X_i]]. \quad (35)$$

With no loss of generality, in this definition we may assume that  $\mathbb{E}f = 0$ .

The following theorem states that the MC ribbon is the largest possible  $\Phi$ -ribbon.

**Theorem 23.** *For any  $\Phi \in \mathcal{F}$  we have  $\mathfrak{R}_\Phi(X_1, \dots, X_k) \subseteq \mathfrak{S}(X_1, \dots, X_k)$ .*

The proof of this theorem is based on Lemma 5 and is similar to that of Theorem 19. So we skip a detailed proof.

### 5.1 Alternative characterizations of the MC ribbon

Next, we discuss alternative characterization of the MC ribbon. We first show that to compute the MC ribbon, it suffices to restrict to a special class of functions  $f(X_{[k]})$ .<sup>1</sup>

**Proposition 24.** *In the definition of the MC ribbon  $\mathfrak{S}(X_1, \dots, X_k)$  we may restrict to functions  $f_{X_{[k]}}$  that of the form  $f(X_{[k]}) = f_1(X_1) + \dots + f_k(X_k)$  where  $f_i$  is a function of  $X_i$  only.*

*Proof.* Let  $\mathcal{F}_{X_{[k]}}$  be the linear space of all functions over  $\mathcal{X}_1 \times \dots \times \mathcal{X}_k$  equipped with the inner product:

$$\langle f, g \rangle := \mathbb{E}[fg].$$

Let  $\mathcal{F}_{X_i} \subseteq \mathcal{F}_{X_{[k]}}$  be the linear space of all functions that depend only on  $X_i$ . Observe that for any  $i \neq j$ ,  $\mathcal{F}_{X_i} \cap \mathcal{F}_{X_j} = \text{span}\{\mathbf{1}_{X_{[k]}}\}$  is the set of constant functions. Moreover,  $\mathcal{F}_{X_i}^0 = \mathcal{F}_{X_i} \cap \mathbf{1}^\perp$  consists of all *zero-mean* functions that depend only on  $X_i$ . Putting these together we have

$$\mathcal{F}_{X_{[k]}} = \text{span}\{\mathbf{1}_{X_{[k]}}\} \oplus \left( \bigoplus_{i=1}^k \mathcal{F}_{X_i}^0 \right) \oplus \mathcal{U}_{X_{[k]}},$$

where

$$\mathcal{U}_{X_{[k]}} = \bigcap_{i=1}^k \mathcal{F}_{X_i}^\perp,$$

is the set of functions that are orthogonal to all functions that depend only on one of  $X_i$ 's. Observe that, for any function  $u_{X_{[k]}} \in \mathcal{U}_{X_{[k]}}$  we have  $\mathbb{E}[u|X_i] = 0$  for all  $i \in [k]$ , because  $\mathbb{E}(\mathbb{E}[u|X_i]^2) = \mathbb{E}(u\mathbb{E}[u|X_i])$  vanishes since it is the inner product between  $u$  and  $\mathbb{E}[u|X_i]$ , which is a function of  $X_i$ .

---

<sup>1</sup> After the MC ribbon was introduced in [3] by the authors, the second author had an email exchange with Sudeep Kamath who claimed the statement of Proposition 24. However, Kamath's proof did not appear rigorous to the second author. Independently, this proposition was found and shown by the first author via a different proof technique.

Let  $f \in \mathcal{F}_{X_{[k]}}$  be an arbitrary function with  $\mathbb{E}f = 0$ . By the above decomposition there exist  $g_i \in \mathcal{F}_{X_i}^0$  and  $u_{X_{[k]}} \in \mathcal{U}_{X_{[k]}}$  such that

$$f_{X_{[k]}} = \mathbb{E}[f] + g_1 + \cdots g_k + u_{X_{[k]}} = g_1 + \cdots g_k + u_{X_{[k]}}.$$

Let  $\tilde{f} = \sum_i g_i$ . Then,  $\mathbb{E}[f|X_i] = \mathbb{E}[\tilde{f}|X_i]$  simply because  $\mathbb{E}[u|X_i] = 0$ . Thus,

$$\begin{aligned} \text{Var}[f] &= \text{Var}[u] + \text{Var}[\tilde{f}], \\ \text{Var}[\mathbb{E}[f|X_i]] &= \text{Var}[\mathbb{E}[\tilde{f}|X_i]]. \end{aligned}$$

Now fixing  $\tilde{f}$ , since  $\text{Var}[u] \geq 0$ , the inequality (35) holds for all  $f = \tilde{f} + u$  if and only if it holds for  $\tilde{f}$ . This completes the proof.  $\square$

We can further simplify calculation of the MC ribbon.

**Theorem 25.** *The MC ribbon  $\mathfrak{S}(X_1, \dots, X_k)$  is equal to the set of  $k$ -tuples  $(\lambda_1, \lambda_2, \dots, \lambda_k) \in [0, 1]^k$  such that for all functions  $f_1(X_1), \dots, f_k(X_k)$ , we have*

$$\text{Var}[f_1 + \cdots + f_k] \leq \sum_{i=1}^k \frac{1}{\lambda_i} \text{Var}[f_i]. \quad (36)$$

Before giving a proof of this theorem, let us discuss some of its implications.

Firstly, computation of the MC ribbon using its characterization given in this theorem is much easier because the dimension of the space of all functions on  $\prod_{i=1}^k \mathcal{X}_i$  quickly becomes very large as we increase  $k$ , the number of variables. However, in the characterization (36) of the MC ribbon we should search on a space of functions whose dimension scales linearly with  $k$ . Moreover, the variance of a conditional expectation in the original definition is replaced by a simple variance that is easier to compute.

Secondly, Equation (36) can be understood as a *strong Cauchy-Schwarz inequality*. Letting  $f_i(X_i)$ 's to be arbitrary functions with zero mean, by the Cauchy-Schwarz inequality we have

$$\mathbb{E}[\sum_{i=1}^k f_i]^2 \leq k \sum_{i=1}^k \mathbb{E}[f_i]^2. \quad (37)$$

Then by Theorem 25, the  $k$ -tuple  $(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k})$  belongs to  $\mathfrak{S}(X_{[k]})$ . Then the MC ribbon characterizes the extent to which this inequality can be strengthened.

Thirdly, from the definition of the MC ribbon it is clear that  $\mathfrak{S}(X_{[k]})$  is convex. Theorem 25 says that the set of  $k$ -tuples  $(\lambda_1^{-1}, \dots, \lambda_k^{-1})$  such that  $\lambda_{[k]} \in \mathfrak{S}(X_{[k]})$ , is convex too. This is a fact that is not clear from the definition of the MC ribbon.

*Proof.* Our proof uses Proposition 24. Consider the space of functions  $\mathcal{F}_{X_i}^0$  of all zero-mean functions of  $X_i$  (as defined in the proof of Proposition 24). With no loss of generality, we can assume that  $p_{X_i}(x_i) > 0$  for all  $x_i$ . Hence,  $\mathcal{F}_{X_i}^0$  is a linear vector space with dimension  $|\mathcal{X}_i| - 1$ . Let  $\{f_{ij} : j = 1, \dots, |\mathcal{X}_i| - 1\}$  be an orthonormal basis for  $\mathcal{F}_{X_i}^0$ . Thus,  $f_{ij}$  is a function of  $X_i$  with zero mean and unit variance, and we have  $\mathbb{E}[f_{ij}f_{ij'}] = 0$  for  $j \neq j'$ . Define

$$m_{i_1j_1; i_2j_2} = m_{i_2j_2; i_1j_1} = \mathbb{E}[f_{i_1j_1}(X_{i_1})f_{i_2j_2}(X_{i_2})].$$

Any arbitrary zero-mean function  $g_i(X_i)$  can be expressed as

$$g_i = \sum_{j=1}^{|\mathcal{X}_i|-1} c_{ij} f_{ij},$$

for some real coefficients  $c_{ij}$ . Then we have  $\text{Var}[g_i] = \sum_j c_{ij}^2$ , and

$$\mathbb{E}[g_{i_1} g_{i_2}] = \sum_{j_1, j_2} c_{i_1 j_1} c_{i_2 j_2} m_{i_1 j_1; i_2 j_2}.$$

Moreover, for any function  $u$  with zero mean,  $\text{Var}[\mathbb{E}[u|X_i]]$ , i.e., the squared length of  $\mathbb{E}[u|X_i]$ , is given by  $\sum_j (\mathbb{E}[u f_{ij}])^2$ . Hence,

$$\text{Var}\left[\mathbb{E}\left[\sum_{i_2=1}^k g_{i_2} \middle| X_{i_1}\right]\right] = \sum_{j_1} \left(\mathbb{E}\left[\sum_{i_2=1}^k g_{i_2} f_{i_1 j_1}\right]\right)^2 \quad (38)$$

$$= \sum_{j_1} \left(\mathbb{E}\left[\sum_{i_2} \sum_{j_2} c_{i_2 j_2} f_{i_2 j_2} f_{i_1 j_1}\right]\right)^2 \quad (39)$$

$$= \sum_{j_1} \left(\sum_{i_2} \sum_{j_2} c_{i_2 j_2} m_{i_1 j_1; i_2 j_2}\right)^2. \quad (40)$$

Putting all these together and using Proposition 24 we find that the MC ribbon is the set of  $k$ -tuples  $\lambda_{[k]}$  such that for all  $c_{ij}$ 's we have

$$\sum_{i_1, i_2, j_1, j_2} c_{i_1 j_1} c_{i_2 j_2} m_{i_1 j_1; i_2 j_2} \geq \sum_{i_1} \lambda_{i_1} \sum_{j_1} \left(\sum_{i_2} \sum_{j_2} c_{i_2 j_2} m_{i_1 j_1; i_2 j_2}\right)^2. \quad (41)$$

Similarly, the ribbon defined by (36) is the set of  $k$ -tuples  $\lambda_{[k]}$  such that for all  $c_{ij}$ 's we have

$$\sum_{i_1, i_2, j_1, j_2} c_{i_1 j_1} c_{i_2 j_2} m_{i_1 j_1; i_2 j_2} \leq \sum_i \frac{1}{\lambda_i} \sum_j c_{ij}^2. \quad (42)$$

Therefore, it remains to show that the ribbons defined by equations (41) and (42) are the same.

Let  $M$  be the matrix whose  $(i_1 j_1, i_2 j_2)$  entry is  $m_{i_1 j_1; i_2 j_2}$ . Note that  $M$  is positive definite since it is a Gram matrix. Also let  $\Lambda$  be the diagonal matrix whose  $(ij, ij)$  entry equals  $\lambda_i$ . Then,  $\lambda_{[k]}$  satisfies (42) for all  $c_{ij}$ 's iff  $M \leq \Lambda^{-1}$ , i.e., iff  $(\Lambda^{-1} - M)$  is positive semidefinite. By the operator monotonicity of the function  $t \mapsto -t^{-1}$ , this is equivalent with  $M^{-1} \geq \Lambda$  as well as  $M \geq M \Lambda M$ . Now a straightforward calculation shows that  $M \geq M \Lambda M$  is equivalent with (41) for all  $c_{ij}$ 's. This completes the proof.  $\square$

Let us go back to the characterization of MC ribbon given in Proposition 24. Let  $f = f_1(X_1) + \dots + f_k(X_k)$  be such that  $f_i(X_i)$  has zero mean for  $i = 1, \dots, k$ . Observe that  $\text{Var}[\mathbb{E}[f|X_i]]$  is the squared length of the projection of  $f$  onto the space of all zero-mean functions of  $X_i$ , namely  $\mathcal{F}_{X_i}^0$ . Then  $\text{Var}[\mathbb{E}[f|X_i]]$  can be bounded from below by the squared of the inner product of  $f$  with some unit vector in  $\mathcal{F}_{X_i}^0$ . In particular, for the choice of

$$\hat{f}_i = \frac{f_i}{\sqrt{\text{Var}[f_i]}}, \quad (43)$$

as a unit vector in  $\mathcal{F}_{X_i}^0$  we have

$$\text{Var}[\mathbb{E}[f|X_i]] \geq \mathbb{E}[f\hat{f}_i]^2.$$

The following theorem shows that using this lower bound in Proposition 24 gives another equivalent representation of the MC ribbon.

**Theorem 26.** *Let  $\mathfrak{S}'(X_1, \dots, X_k)$  be the set of the  $k$ -tuples  $(\lambda_1, \dots, \lambda_k) \in [0, 1]^k$  such that for all functions  $f_i(X_i)$ ,  $i = 1, \dots, k$ , we have*

$$\text{Var}[f] \geq \sum_{i=1}^k \lambda_i \mathbb{E}[f\hat{f}_i]^2,$$

where  $f = \sum_{i=1}^k f_i$  and  $\hat{f}_i$  is given by (43). Then we have  $\mathfrak{S}'(X_{[k]}) = \mathfrak{S}(X_{[k]})$ .

The proof of this theorem is based on similar ideas as in the proof of Theorem 25, so we leave it for Appendix D.

## 5.2 Extreme MC ribbons

It is well-known that  $\rho(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent. The following proposition is a generalization of this fact.

**Proposition 27.**  *$\mathfrak{S}(X_1, \dots, X_k)$  is equal to  $[0, 1]^k$  if and only if  $X_1, X_2, \dots, X_k$  are pairwise independent.*

Note that the HC ribbon is equal to  $[0, 1]^k$  if  $X_i$ 's are *mutually* independent. Thus, MC ribbon and HC ribbon behave completely differently and provide different characterizations of the correlations in  $(X_1, \dots, X_k)$  when  $k \geq 3$ .

The tensorization property of the MC ribbon implies that

$$\mathfrak{S}(X_1, X_2, \dots, X_k) = \mathfrak{S}(M_1 X_1, M_2 X_2, \dots, M_k X_k)$$

if  $M_i$ s are independent of  $X_i$ s, and  $M_i$ s are pairwise independent. This fact shows that we can prove infeasibility for the non-interactive distribution simulation problem [1], in the presence of “private” randomnesses  $M_j$  that are pairwise (and not mutually independent).

*Proof.* We may use Theorem 25. If  $X_i$ 's are pairwise independent we clearly have  $\text{Var}[f_1 + \dots + f_k] = \text{Var}[f_1] + \dots + \text{Var}[f_k]$  and then  $\mathfrak{S}(X_{[k]}) = [0, 1]^k$ . Conversely, suppose that  $\text{Var}[f_1 + \dots + f_k] \leq \text{Var}[f_1] + \dots + \text{Var}[f_k]$ , for all  $f_i(X_i)$ 's. By letting  $f_\ell$ 's to be equal to the zero function except the  $i$ -th and  $j$ -th ones, we find that  $\text{Var}[f_i + f_j] \leq \text{Var}[f_i] + \text{Var}[f_j]$  for all  $f_i(X_i)$  and  $f_j(X_j)$ . This means that  $X_i$  and  $X_j$  are independent.  $\square$

Recall that  $\mathfrak{S}(X_1, \dots, X_k)$  always contains  $\{\lambda_{[k]} | \lambda_i \geq 0, \sum_i \lambda_i \leq 1\}$ . The following proposition characterizes the other extreme for the MC ribbon.

**Proposition 28.**  *$\mathfrak{S}(X_{[k]}) = \{\lambda_{[k]} | \lambda_i \geq 0, \sum_i \lambda_i \leq 1\}$  if and only if  $X_i$ 's have a common part, i.e., there are non-constant functions  $g_i(X_i)$ ,  $i = 1, \dots, k$ , such that  $g_1(X_1) = \dots = g_k(X_k)$  with probability one.*

*Proof.* If  $X_i$ 's have a common part, for  $f(X_{[k]}) = g_1(X_1) = \dots = g_k(X_k)$  we have  $\text{Var}[f] = \text{Var}[\mathbb{E}[f|X_i]]$ . Therefore, by the definition of the MC ribbon we have  $\mathfrak{S}(X_{[k]}) = \{\lambda_{[k]} | \lambda_i \geq 0, \sum_i \lambda_i \leq 1\}$ .

Conversely, assume that  $\mathfrak{S}(X_{[k]}) = \{\lambda_{[k]} : \lambda_i \geq 0, \sum_i \lambda_i \leq 1\}$ . Then, the maximum value of  $\lambda$  such that  $(\lambda, \lambda, \dots, \lambda)$  is in  $\mathfrak{S}(X_{[k]})$ , is equal to  $1/k$ . This maximum value of  $\lambda$  can be written as

$$\lambda_{\max} = \inf_f \frac{\text{Var}[f]}{\sum_{i=1}^k \text{Var}[\mathbb{E}[f|X_i]]}.$$

Observe that by scaling  $f$ , we can restrict the infimum to functions satisfying  $\text{Var}[f] = 1$ . Then by a compactness argument, the infimum is achieved at some (non-constant) function  $f$ . For this function we have  $\sum_{i=1}^k \text{Var}[\mathbb{E}[f|X_i]] = k \text{Var}[f]$ . This means that all the inequalities  $\text{Var}[\mathbb{E}[f|X_i]] \leq \text{Var}[f]$  are equality for  $f$ . That is, by the law of total variance, we have  $\text{Var}[f|X_i] = 0$ . In other words,  $f$  is a function of  $X_i$  for all  $i \in [k]$ , and a common part. □

### 5.3 Examples

We now compute the MC ribbon for some examples of  $(X_1, \dots, X_k)$ . We first focus on the bipartite case.

**Proposition 29.** *For  $k = 2$  we have*

$$\mathfrak{S}(X_1, X_2) = \left\{ (\lambda_1, \lambda_2) \in [0, 1]^2 \mid \left(1 - \frac{1}{\lambda_1}\right) \left(1 - \frac{1}{\lambda_2}\right) \geq \rho(X_1, X_2)^2 \right\}. \quad (44)$$

It was proved in [3] that the right hand side in (44) always contains  $\mathfrak{S}(X_1, X_2)$ . Here we prove that indeed equality holds.

*Proof.* Using Theorem 25,  $\mathfrak{S}(X_1, X_2)$  is equal to the set of pairs  $(\lambda_1, \lambda_2)$  such that

$$\text{Var}[f_1 + f_2] \leq \frac{1}{\lambda_1} \text{Var}[f_1] + \frac{1}{\lambda_2} \text{Var}[f_2],$$

for all zero-mean functions  $f_1(X_1)$  and  $f_2(X_2)$ . This inequality is equivalent to

$$2\mathbb{E}[f_1 f_2] \leq \left(1 - \frac{1}{\lambda_1}\right) \text{Var}[f_1] + \left(1 - \frac{1}{\lambda_2}\right) \text{Var}[f_2].$$

Then the desired result follows once we note that

$$\rho(X, Y) = \max \frac{\mathbb{E}[g_1 g_2]}{\sqrt{\text{Var}[g_1] \text{Var}[g_2]}},$$

where the maximum is over all zero-mean functions  $g_1(X_1)$  and  $g_2(X_2)$ . □

Let us now consider computing the MC ribbon for multivariate distributions, i.e., for  $k \geq 3$ . Observe that if  $X_i$  is binary (taking values in a binary set), then there is a unique (up to a constant) function  $f_i(X_i)$  that has zero mean. Then computing the MC ribbon using Theorem 25 is not hard if  $X_i$ 's are all binary. See Theorem 31 for details.

**Binary-Binary-Ternary:** Assume that  $k = 3$ , and that  $X_1, X_2$  are binary, and  $X_3$  is ternary. Let  $\rho_{ij} = \rho(X_i, X_j)$  be the maximal correlation coefficient between  $X_i$  and  $X_j$  for distinct  $i, j \in \{1, 2, 3\}$ . Since  $X_1$  and  $X_2$  are binary, there are unique (up to a sign) zero-mean functions  $g_1(X_1)$  and  $g_2(X_2)$  with unit variance. We choose the sign of such functions  $g_1(X_1)$  and  $g_2(X_2)$  such that

$$\mathbb{E}[g_1 g_2] = \rho_{12}. \quad (45)$$

Again by the uniqueness of  $g_i$ ,  $i = 1, 2$ , and that  $\eta_\Psi(X_i, X_3) = \rho^2(X_i, X_3)$ , for  $\Psi(t) = t^2$ , we have

$$\rho_{13}^2 = \text{Var}[\mathbb{E}[g_1|X_3]], \quad (46)$$

and

$$\rho_{23}^2 = \text{Var}[\mathbb{E}[g_2|X_3]]. \quad (47)$$

Finally define

$$r_{12 \rightarrow 3} = \mathbb{E}[\mathbb{E}[g_1|X_3] \mathbb{E}[g_2|X_3]]. \quad (48)$$

Assume that the distribution of  $(X_1, X_2, X_3)$  is generic, so that the functions  $\mathbb{E}[g_1|X_3]$  and  $\mathbb{E}[g_2|X_3]$  are linearly independent.

**Proposition 30.** *Under the assumptions given above,  $\mathfrak{S}(X_1, X_2, X_3)$  is the set of triples  $(\lambda_1, \lambda_2, \lambda_3) \in [0, 1]^3$  such that the followings hold:*

$$\begin{aligned} \left(\frac{1}{\lambda_1} - 1\right)\left(\frac{1}{\lambda_3} - 1\right) &\geq \rho_{13}^2, \\ \left(\frac{1}{\lambda_2} - 1\right)\left(\frac{1}{\lambda_3} - 1\right) &\geq \rho_{23}^2, \\ \left(\left(\frac{1}{\lambda_1} - 1\right)\left(\frac{1}{\lambda_3} - 1\right) - \rho_{13}^2\right)\left(\left(\frac{1}{\lambda_2} - 1\right)\left(\frac{1}{\lambda_3} - 1\right) - \rho_{23}^2\right) &\geq \left(\left(\frac{1}{\lambda_3} - 1\right)\rho_{12} + r_{12 \rightarrow 3}\right)^2. \end{aligned}$$

Observe that by the above theorem, the MC ribbon of  $(X_1, X_2, X_3)$  cannot be computed solely based on the marginal distributions of pairwise random variables (compare this with Proposition 27).

The proof of this proposition is given in Appendix E.

**Normal distributions:** Throughout the paper, we considered only discrete random variables, i.e.,  $\mathcal{X}_i$ 's are finite sets. Nevertheless, the definition of  $\Phi$ -ribbon can easily be generalized to the continuous case. Moreover, most of the properties that we proved here, except those that are based on compactness, are generalized for continuous random variables as well. Here, our goal is to compute the MC ribbon for multivariate normal distributions using Theorem 25. Since we have not presented the proof of this theorem in the continuous case, the reader may consider the statement of Theorem 25 as the definition of MC ribbon for normal distributions.

Let  $(X_1, \dots, X_k)$  be *real* random variables that are either binary (i.e., the alphabet set of  $X_i$  is of size two), or normal (i.e.,  $X_i$ 's form a multivariate normal distribution). Let  $R$  be the *covariance matrix* of  $X_i$ 's. That is,  $R$  is a matrix whose  $(i, j)$ -th entry is the Pearson correlation coefficient between  $X_i$  and  $X_j$ :

$$R_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}[X_i]\text{Var}[X_j]}},$$

where  $\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$ . Next, given a  $k$ -tuple  $(\lambda_1, \lambda_2, \dots, \lambda_k) \in \mathfrak{S}$ , we associate to it a diagonal matrix  $\Lambda$  whose  $i$ -th entry on the diagonal is equal to  $\lambda_i$ .



**Theorem 31.** Suppose that  $(X_1, \dots, X_k)$  either form a multivariate normal distribution, or are all binary taking values in an alphabet set of size two. Let  $R$  be the covariance matrix of  $(X_1, \dots, X_k)$  as defined above. Then,  $(\lambda_1, \dots, \lambda_k)$  belongs to  $\mathfrak{S}(X_1, \dots, X_k)$  if and only if  $R \leq \Lambda^{-1}$ .

The proof of this theorem is based on ideas from [15] and is given in Appendix F.

## 5.4 Another multipartite correlation region

Motivated by the form of characterization of  $\mathfrak{S}(X_1, \dots, X_k)$  given by Theorem 25, we define another region associate to a multivariate distribution.

**Definition 32.** For any  $(X_1, \dots, X_k)$  we define  $\tilde{\mathfrak{S}}(X_1, \dots, X_k)$  to be the set of  $k$ -tuples  $(\lambda_1, \dots, \lambda_k) \in [0, 1]^k$  such that for all functions  $f_i(X_i)$ ,  $i = 1, \dots, k$ , we have

$$\text{Var}[f_1 + \dots + f_k] \geq \sum_{i=1}^k \lambda_i \text{Var}[f_i].$$

$\tilde{\mathfrak{S}}(X_1, \dots, X_k)$  and the MC ribbon share the properties of monotonicity and tensorization, yet as we will argue later, they are not identical.

**Theorem 33.**  $\tilde{\mathfrak{S}}(X_1, \dots, X_k)$  satisfies the monotonicity and tensorization properties the same as  $\mathfrak{S}(X_1, \dots, X_k)$ .

This theorem is proved in Appendix G.

**Proposition 34.**  $\tilde{\mathfrak{S}}(X_1, \dots, X_k)$  is equal to  $[0, 1]^k$  if and only if  $X_1, X_2, \dots, X_k$  are pairwise independent.

The proof of the following proposition is similar to that of Proposition 27, so we do not repeat it here.

The above proposition characterizes one extreme of  $\tilde{\mathfrak{S}}(X_1, \dots, X_k)$  that is common with the MC ribbon. To characterize the other extreme of  $\tilde{\mathfrak{S}}(X_1, \dots, X_k)$ , recall that the MC ribbon contains all  $\lambda_{[k]} \in [0, 1]^k$  with  $\sum_i \lambda_i \leq 1$ . Nevertheless, these points may not belong to  $\tilde{\mathfrak{S}}(X_1, \dots, X_k)$ . Indeed, we can even have  $\tilde{\mathfrak{S}}(X_{[k]}) = \{(0, 0, \dots, 0)\}$ . To see this, let  $k = 2$  and assume that  $X_1 = X_2$  with probability one. Let  $f_1 = -f_2$  be a non-constant function of both  $X_1 = X_2$ . Then,  $f_1 + f_2 = 0$ , and thus  $\text{Var}[f_1 + f_2] = 0$ . But  $\text{Var}[f_1]$  and  $\text{Var}[f_2]$  are both positive, so  $\tilde{\mathfrak{S}}(X_1, X_2)$  contains only  $(0, 0)$ .

**Theorem 35.**  $\tilde{\mathfrak{S}}(X_{[k]}) = \{(0, \dots, 0)\}$  if and only if there are zero-mean functions  $f_1(X_1), \dots, f_k(X_k)$  that are not all zero and  $f_1 + \dots + f_k = 0$ .

The proof of this theorem is given in Appendix H.

Based on Theorem 35, for  $k = 2$ ,  $\tilde{\mathfrak{S}}(X_1, X_2) = \{(0, 0)\}$  if and only if  $X_1$  and  $X_2$  have common data. However, for  $k \geq 3$  one can find examples of  $(X_1, \dots, X_k)$  that do not have common part, yet we have  $\tilde{\mathfrak{S}}(X_{[k]}) = \{(0, \dots, 0)\}$ .

**Example 36.** Let  $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$ . Let  $0 < a, b < 1$  be such that  $c = a + b < 1$ . Define  $p(x, y, z)$  by

$$p(000) = a, \quad p(110) = b, \quad p(101) = 1 - c,$$

and  $p(xyz) = 0$  for  $(x, y, z) \notin \{(0, 0, 0), (1, 1, 0), (1, 0, 1)\}$ . Observe that  $X, Y, Z$  do not have common data. Now define  $f_X, g_Y, h_Z$  as follows:

$$\begin{aligned} f_X(0) &= 1 - a, & f_X(1) &= a, \\ g_Y(0) &= -b, & g_Y(1) &= 1 - b, \\ h_Z(0) &= c - 1, & h_Z(1) &= c. \end{aligned}$$

Then, we have  $\mathbb{E}[f_X] = \mathbb{E}[g_Y] = \mathbb{E}[h_Z] = 0$ ,  $\mathbb{E}[f_X^2] > 0$ ,  $\mathbb{E}[g_Y^2] > 0$ ,  $\mathbb{E}[h_Z^2] > 0$  and  $f(X) + g(Y) + h(Z) = 0$ . Therefore,  $\tilde{\mathfrak{S}}(X, Y, Z) = \{(0, 0, 0)\}$ .

In the following theorem we use the same notation as we used in Theorem 31. The proof of this theorem is also similar to that of Theorem 31, so we do not repeat it here.

**Theorem 37.** Suppose that  $(X_1, \dots, X_k)$  either form a multivariate normal distribution, or are all binary  $|\mathcal{X}_i| = 2$ . Let  $R$  be the covariance matrix of  $(X_1, \dots, X_k)$  as defined above. Then,  $(\lambda_1, \dots, \lambda_k)$  belongs to  $\mathfrak{S}(X_1, \dots, X_k)$  if and only if  $R \leq \Lambda$ .

## 6 Summary of the results

In this paper, we defined  $\Phi$ -ribbon, that generalizes both the MC and the HC ribbons. We showed that the  $\Phi$ -ribbon satisfies the monotonicity and tensorization properties. Therefore,  $\Phi$ -entropy can be utilized in any of the known applications of the HC ribbon and the maximal correlation in network information theory, namely the non-interactive distribution simulation problem [1] or transmission of correlated sources over a noisy network (such as a MAC channel) [2]. It was shown that the  $\Phi$ -ribbon relates to Raginsky's  $\Phi$ -strong data processing constant. Next, we showed that the MC ribbon is the maximal  $\Phi$ -ribbon, *i.e.*, all  $\Phi$ -ribbons are subsets of the MC ribbon. This fact motivated further study of the properties of the MC ribbon, and its efficient calculation. In particular, an equivalent characterization of the MC ribbon was given. Inspired by the form of this characterization, we defined another multivariate correlation region that is characterized by maximal correlation when  $k = 2$ , and satisfies the data processing and tensorization properties.

## References

- [1] S. Kamath and V. Anantharam, "Non-interactive Simulation of Joint Distributions: The Hirschfeld-Gebelein-Rényi Maximal Correlation and the Hypercontractivity Ribbon," Proceedings of the 50th Annual Allerton Conference on Communications, Control and Computing (2012).
- [2] W. Kang and S. Ulukus, "A New Data Processing Inequality and Its Applications in Distributed Source and Channel Coding," IEEE Transactions on Information Theory, 57 (1): 56-69, 2011.
- [3] S. Beigi and A. Gohari, "Monotone Measures for Non-Local Correlations," IEEE Transactions on Information Theory, 61 (9): 5185-5208, 2015.
- [4] H. O. Hirschfeld, "A connection between correlation and contingency," Proc. Cambridge Philosophical Soc. **31**, 520-524 (1935).
- [5] H. Gebelein, "Das statistische problem der Korrelation als variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichungsrechnung," Z. für angewandte Math. und Mech. **21**, 364-379 (1941).

- [6] A. Rényi, “New version of the probabilistic generalization of the large sieve,” *Acta Math. Hung.* **10**, 217-226 (1959).
- [7] A. Rényi, “On measures of dependence,” *Acta Math. Hung.* **10**, 441-451 (1959).
- [8] H. S. Witsenhausen, “On sequences of pairs of dependent random variables,” *SIAM Journal on Applied Mathematics*, 28 (1): 100-113, 1975.
- [9] P. Gács and J. Körner, “Common information is far less than mutual information,” *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 119-162, 1972.
- [10] S. M. Ali and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another,” *Journal of the Royal Statistics Society, series B*, vol. 28, no. 1, pp. 131-142, 1966.
- [11] I. Csiszar, “Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markhoffschen Ketten,” *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85-108, Jan. 1963.
- [12] I. Csiszar, “Information-type measures of difference of probability distributions and indirect observations,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299-318, Jan. 1967.
- [13] A. Dembo, A. Kagan, and L. Shepp, “Remarks on the maximum correlation coefficient,” *Bernoulli*, vol. 7, pp. 343-350, 2001.
- [14] S. Kamath and C. Nair, “The strong data processing constant for sums of i.i.d. random variables,” *Information Theory (ISIT)*, 2015 IEEE International Symposium on, pp. 2550-2552.
- [15] H. O. Lancaster, “Some properties of the bivariate normal distribution considered in the form of a contingency table,” *Biometrika*, 44, 1-2: 289-292 (1957).
- [16] C. Nair, “Equivalent formulations of Hypercontractivity using Information Measures,” IZS workshop, 2014, available at <http://chandra.ie.cuhk.edu.hk/pub/papers/manuscripts/IZS14.pdf>
- [17] R. Ahlswede and P. Gács, “Spreading of Sets in Product Spaces and Hypercontraction of the Markov Operator,” *The Annals of Probability* 4, 925-939 (1976).
- [18] E. A. Carlen and C. E. Dario, “Subadditivity of the entropy and its relation to Brascamp-Lieb type inequalities,” *Geometric and Functional Analysis*, vol. 19, no. 2, pp. 373-405, 2009.
- [19] M. Raginsky, “Strong Data Processing Inequalities and  $\Phi$ -Sobolev Inequalities for Discrete Channels,” *IEEE Transactions on Information Theory*, 62 (6), 3355-3389, 2016.
- [20] V. Anantharam, A. Gohari, S. Kamath, C. Nair, “On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover,” *arXiv preprint arXiv:1304.6133*, 2013.
- [21] R. Bhatia, *Positive definite matrices*. Princeton university press, 2009.
- [22] S. Boucheron, C. Lugosi, P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [23] D. Chafaï, Entropies, convexity, and functional inequalities, On  $\Phi$ -entropies and  $\Phi$ -Sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44 (2), 325-363, 2004.

[24] D. Chafaï, Binomial-Poisson entropic inequalities and the  $M/M/\infty$  queue. ESAIM: Probability and Statistics, 10, 317-339, 2006.

## A SDPI constant

In this appendix we prove Theorem 21 as well as the claim we made in Example 17. Let us start with the latter.

**Proposition 38.** *If  $(X, Y)$  are distributed according to  $\text{DSBS}(\lambda)$ , then for any  $\Phi \in \mathcal{F}$  we have*

$$\eta_\Phi(X, Y) = \lambda^2.$$

*Proof.* By Theorem 19 we know that  $\eta_\Phi(X, Y) \geq \rho^2(X, Y) = \lambda^2$ . To prove the inequality in the other direction we need to show that for any function  $f_X$  we have

$$\lambda^2 H_\Phi(f) \geq H_\Phi(\mathbb{E}[f|Y]).$$

Let  $m = \mathbb{E}f$ , and define  $z$  by  $f(0) = m + z$ . Then  $f(1) = m - z$ , and the above inequality reduced to

$$\lambda^2 (\Phi(m + z) + \Phi(m - z) - 2\Phi(m)) \geq \Phi(m + \rho z) + \Phi(m - \rho z) - 2\Phi(m).$$

Let us for  $t \geq 0$  define

$$\psi(t) = \Phi(m + \sqrt{t}) + \Phi(m - \sqrt{t}) - 2\Phi(m).$$

Then the above inequality is equivalent to

$$\lambda^2 \psi(t^2) \geq \psi(\lambda^2 t^2).$$

Since  $\psi(0) = 0$ , this inequality is proven once we show that  $\psi$  is convex. We compute

$$\psi'(t) = \frac{1}{2\sqrt{t}} \Phi'(m + \sqrt{t}) - \frac{1}{2\sqrt{t}} \Phi'(m - \sqrt{t}),$$

and

$$\psi''(t) = -\frac{1}{4t^{3/2}} (\Phi'(m + \sqrt{t}) - \Phi'(m - \sqrt{t})) + \frac{1}{4t} (\Phi''(m + \sqrt{t}) + \Phi''(m - \sqrt{t})).$$

Then the convexity of  $\psi(x)$  is equivalent to

$$\Phi''(m + \sqrt{t}) + \Phi''(m - \sqrt{t}) \geq \frac{1}{\sqrt{t}} (\Phi'(m + \sqrt{t}) - \Phi'(m - \sqrt{t})).$$

Equivalently we need to show that for any  $s \geq 0$  we have

$$\Phi''(m + s) + \Phi''(m - s) \geq \frac{1}{s} (\Phi'(m + s) - \Phi'(m - s)).$$

Define  $\xi(s) = \Phi'(m + s) - \Phi'(m - s)$ . Then the above inequality can be rewritten as

$$\xi'(s) \geq \frac{1}{s} (\xi(s) - \xi(0)),$$

which holds if  $\xi'(s)$  is increasing since  $s \geq 0$ . Equivalently we need to prove  $\xi''(s) \geq 0$  for all  $s \geq 0$ . That is, we want

$$\Phi'''(m + s) \geq \Phi'''(m - s),$$

which holds if  $\Phi'''$  is increasing.

By part (vi) of the definition of class of functions  $\mathcal{F}$ , we have  $\Phi''''\Phi'' \geq 2\Phi'''^2$ . On the other hand,  $\Phi$  is convex which means that  $\Phi'' \geq 0$ . Then by this inequality we have  $\Phi''' \geq 0$ . As a result,  $\Phi'''$  is increasing. We are done.  $\square$

Before proving Theorem 21 let us first generalize the definition of the  $\Phi$ -SDPI constant.

**Definition 39.** For any pair of convex functions  $\Phi, \Psi$  we define  $\eta_{\Phi, \Psi}(X, Y)$ , to be the smallest (the infimum of)  $\lambda \geq 0$  such that for any function  $f_X$  we have

$$\lambda H_{\Phi}(f) \geq H_{\Psi}(\mathbb{E}[f|Y]).$$

Observe that  $\eta_{\Phi, \Psi}(X, Y)$  may be greater than one for arbitrary  $\Phi$  and  $\Psi$ . Moreover,  $\eta_{\Phi, \Psi}(X, Y)$  coincides with  $\eta_{\Phi}(X, Y)$  when  $\Psi = \Phi$ .

The first property that  $\eta_{\Phi, \Psi}$  share with the  $\Phi$ -SDPI constant is Proposition 18, that for a fixed  $p_X$  the function

$$p_{Y|X} \mapsto \eta_{\Phi, \Psi}(X, Y),$$

is convex. The proof of this fact is again based on Lemma 4.

**Theorem 40.** Let  $\Phi$  and  $\Psi$  be convex functions, and assume that at least one of them belongs to  $\mathcal{F}$ . Then the followings hold.

(i) *Monotonicity:* If  $X - A - B - Y$  then  $\eta_{\Phi, \Psi}(X, Y) \leq \eta_{\Phi, \Psi}(A, B)$ .

(ii) *Tensorization:* If  $p_{ABXY} = p_{AB} \cdot p_{XY}$  then

$$\eta_{\Phi, \Psi}(AX, BY) = \max \{ \eta_{\Phi, \Psi}(A, B), \eta_{\Phi, \Psi}(X, Y) \}.$$

The proof of is theorem is very similar to that of Theorem 12.

*Proof.* (i) Let  $f_X$  be a function of  $X$ . Then  $\mathbb{E}[f|A]$  is a function of  $A$ , and by the definition of  $\eta_{\Phi, \Psi}(A, B)$  we have

$$\eta_{\Phi, \Psi}(A, B) H_{\Phi}(f) \geq \eta_{\Phi, \Psi}(A, B) H_{\Phi}(\mathbb{E}[f|A]) \geq H_{\Psi}(\mathbb{E}[\mathbb{E}[f|A]|B]) = H_{\Psi}(\mathbb{E}[f|B]),$$

where the equality follows since we have the Markov chain  $X - A - B$ . Letting  $g_B = \mathbb{E}[f|B]$  we have

$$H_{\Psi}(g) \geq H_{\Psi}(\mathbb{E}[g|Y]) = H_{\Psi}(\mathbb{E}[f|Y]),$$

since  $X - B - Y$  forms a Markov chain. Putting these together we arrive at  $\eta_{\Phi, \Psi}(A, B) H_{\Phi}(f) \geq H_{\Psi}(\mathbb{E}[f|Y])$ . As a result,  $\eta_{\Phi, \Psi}(X, Y) \leq \eta_{\Phi, \Psi}(A, B)$ .

(ii) By restricting to functions that depend only on  $A$  or on  $X$ , it is easy to see that

$$\eta_{\Phi, \Psi}(AX, BY) \geq \max \{ \eta_{\Phi, \Psi}(A, B), \eta_{\Phi, \Psi}(X, Y) \}.$$

Let  $\lambda = \max \{ \eta_{\Phi, \Psi}(A, B), \eta_{\Phi, \Psi}(X, Y) \}$ . Then we need to show that  $\lambda \geq \eta_{\Phi, \Psi}(AX, BY)$ . To prove this we need show that for any function  $f_{AX}$  we have

$$\lambda H_{\Phi}(f) \geq H_{\Psi}(\mathbb{E}[f|BY]). \quad (49)$$

First assume that  $\Psi \in \mathcal{F}$ . Since  $\lambda \geq \eta_{\Phi, \Psi}(X, Y)$  and the distribution of  $(X, Y)$  does not change when we condition on  $A$ , we have

$$\lambda H_{\Phi}(f|A) \geq H_{\Psi}(\mathbb{E}[f|AY]|A).$$

On the other hand since  $\lambda \geq \eta_{\Phi, \Psi}(A, B)$ , for  $\mathbb{E}[f|A]$ , as a function of  $A$ , we have

$$\lambda H_{\Phi}(\mathbb{E}[f|A]) \geq H_{\Psi}(\mathbb{E}[\mathbb{E}[f|A]|B]) = H_{\Psi}(\mathbb{E}[f|B]),$$

where the equality follows from the independence of  $X$  and  $(A, B)$ . Summing up the above two inequalities we obtain

$$\lambda H_{\Phi}(f) \geq H_{\Psi}(\mathbb{E}[f|AY]|A) + H_{\Psi}(\mathbb{E}[f|B]).$$

Then it suffices to show that

$$H_{\Psi}(\mathbb{E}[f|AY]|A) + H_{\Psi}(\mathbb{E}[f|B]) \geq H_{\Psi}(\mathbb{E}[f|BY]).$$

Let  $g_{AY} = \mathbb{E}[f|AY]$ . Then we have  $\mathbb{E}[f|B] = \mathbb{E}[g|B]$  and  $\mathbb{E}[g|BY] = \mathbb{E}[f|BY]$ . Therefore, this inequality is equivalent to

$$H_{\Psi}(g|A) + H_{\Psi}(\mathbb{E}[g|B]) \geq H_{\Psi}(\mathbb{E}[g|BY]).$$

Using  $\Psi \in \mathcal{F}$ , this inequality follows from part (c) of Lemma 7 and  $H_{\Psi}(g|A) \geq H_{\Psi}(g|AB)$ .

Now we assume that  $\Phi \in \mathcal{F}$  and prove (49). Since  $\lambda \geq \eta_{\Phi, \Psi}(X, Y)$ , and the distribution of  $(X, Y)$  does not change when we condition on  $B$ , we have

$$\lambda H_{\Phi}(\mathbb{E}[f|XB]|B) \geq H_{\Psi}(\mathbb{E}[\mathbb{E}[f|XB]|YB]|B) = H_{\Psi}(\mathbb{E}[f|BY]|B),$$

Similarly, since  $\lambda \geq \eta_{\Phi, \Psi}(A, B)$ , for  $\mathbb{E}[f|A]$  as function of  $A$  we have

$$\lambda H_{\Phi}(\mathbb{E}[f|A]) \geq H_{\Psi}(\mathbb{E}[\mathbb{E}[f|A]|B]) = H_{\Psi}(\mathbb{E}[f|B]) = H_{\Psi}(\mathbb{E}[\mathbb{E}[f|BY]|B]).$$

Summing up the above two inequalities and using the chain rule we find that

$$\lambda (H_{\Phi}(\mathbb{E}[f|XB]|B) + H_{\Phi}(\mathbb{E}[f|A])) \geq H_{\Psi}(\mathbb{E}[f|BY]).$$

Then it suffices to show that

$$H_{\Phi}(f) \geq H_{\Phi}(\mathbb{E}[f|XB]|B) + H_{\Phi}(\mathbb{E}[f|A]).$$

Observe that since  $\Phi \in \mathcal{F}$  and  $A - B - X$  forms a Markov chain, by part (b) of Lemma 7 we have  $H_{\Phi}(\mathbb{E}[f|XB]|B) \leq H_{\Phi}(f|AB) \leq H_{\Phi}(f|A)$ . The above inequality then follows from the chain rule.  $\square$

Note that in the proof of the monotonicity property we use only the convexity of  $\Phi$  and  $\Psi$ .

## B Proof of Theorem 14

Given  $u \in \mathcal{U}$ , let

$$f_u(x_{[k]}) = \frac{p(x_{[k]}, u)}{p(x_{[k]})p(u)},$$

Then,  $f_u \geq 0$  and  $\mathbb{E}_{p(x_{[k]})}[f_u(X_{[k]})] = 1$ , and

$$\sum_u p(u) H_\Phi(f_u) = I_\Phi(X_{[k]}; U)$$

Furthermore,

$$\mathbb{E}[f_u(X_{[k]})|X_i = x_i] = \sum_{x_{[i]}} p(x_{[i]}|x_i) f_u(x_{[i]}) \quad (50)$$

$$= \sum_{x_{[i]}} p(x_{[i]}|x_i) \frac{p(x_{[k]}, u)}{p(x_{[k]})p(u)} \quad (51)$$

$$= \frac{p(x_i, u)}{p(x_i)p(u)} \quad (52)$$

As a result

$$\sum_u p(u) H_\Phi(\mathbb{E}[f_u|X_i]) = I_\Phi(X_i; U).$$

Now, if  $\lambda_{[k]} \in \mathfrak{R}'_\Phi(X_{[k]})$ , for any  $u$  we have

$$\sum_i \lambda_i H_\Phi(\mathbb{E}[f_u|X_i]) \leq H_\Phi(f_u).$$

Multiplying the above in  $p(u)$  and adding up over  $u$ , we get that  $\lambda_{[k]}$  is in the  $\mathfrak{R}_{I_\Phi}(X_{[k]})$ . Hence,  $\mathfrak{R}'_\Phi(X_{[k]}) \subseteq \mathfrak{R}_{I_\Phi}(X_{[k]})$ .

Conversely, to show that  $\mathfrak{R}_{I_\Phi}(X_{[k]}) \subseteq \mathfrak{R}'_\Phi(X_{[k]})$  we adopt the construction from [20]. Assume that  $\lambda_{[k]}$  is in  $\mathfrak{R}_{I_\Phi}(X_{[k]})$ . Let  $f(x_1, \dots, x_k)$  be an arbitrary function satisfying  $f \geq 0$  and  $\mathbb{E}[f] = 1$ . We want to show that

$$\sum_i \lambda_i H_\Phi(\mathbb{E}[f|X_i]) \leq H_\Phi(f).$$

Without loss of generality we may assume that  $f(x_1, \dots, x_n)$  is zero whenever  $p(x_1 \dots x_n)$  is zero. Define the distribution  $p_\epsilon(u, x_1, \dots, x_k)$  as follows. Let  $U_\epsilon$  be a binary random variable such that  $p_\epsilon(U_\epsilon = 0) = \epsilon$  and  $p_\epsilon(U_\epsilon = 1) = 1 - \epsilon$ . Also let

$$\begin{aligned} p_\epsilon(x_1, \dots, x_k | U_\epsilon = 0) &= p(x_1, \dots, x_k) f(x_1, \dots, x_k) \\ p_\epsilon(x_1, \dots, x_k | U_\epsilon = 1) &= p(x_1, \dots, x_k) \left( \frac{1}{1 - \epsilon} - \frac{\epsilon}{1 - \epsilon} f(x_1, \dots, x_k) \right) \end{aligned}$$

Observe that for sufficiently small  $\epsilon \geq 0$ ,  $p_\epsilon(u, x_1, \dots, x_k)$  is a probability distribution because  $f \geq 0, \mathbb{E}[f] = 1$ . Moreover, we have  $p_\epsilon(x_1, \dots, x_k) = p(x_1, \dots, x_k)$ . Then since  $(\lambda_1, \dots, \lambda_k)$  is in  $I_\Phi$ -ribbon, we have

$$\sum_i \lambda_i I_\Phi(U_\epsilon; X_i) \leq I_\Phi(U_\epsilon; X_1 \dots X_k).$$

Indeed the function

$$t(\epsilon) = I_\Phi(U_\epsilon; X_1 \dots X_k) - \sum_i \lambda_i I_\Phi(U_\epsilon; X_i)$$

is non-negative for sufficiently small  $|\epsilon|$ . On the other hand, we have  $t(0) = 0$ . Then we should have  $t'(0) \geq 0$ . Observe that

$$I_{\Phi}(U_{\epsilon}; X_1 \dots X_k) = \sum_{x_{[k]}} \epsilon p(x_{[k]}) \Phi\left(\frac{\epsilon p(x_{[k]}) f(x_{[k]})}{\epsilon p(x_{[k]})}\right) + \sum_{x_{[k]}} (1 - \epsilon) p(x_{[k]}) \Phi\left(\frac{p(x_{[k]}) (1 - \epsilon f(x_{[k]}))}{(1 - \epsilon) p(x_{[k]})}\right) \quad (53)$$

$$= \sum_{x_{[k]}} \epsilon p(x_{[k]}) \Phi(f(x_{[k]})) + \sum_{x_{[k]}} (1 - \epsilon) p(x_{[k]}) \Phi\left(\frac{1 - \epsilon f(x_{[k]})}{1 - \epsilon}\right) \quad (54)$$

Then,

$$\left. \frac{\partial}{\partial \epsilon} I_{\Phi}(U_{\epsilon}; X_1 \dots X_k) \right|_{\epsilon=0} = \sum_{x_{[k]}} p(x_{[k]}) \Phi(f(x_{[k]})) - \Phi(1) \quad (55)$$

$$= H_{\Phi}(f) \quad (56)$$

Now, observe that

$$\begin{aligned} p_{\epsilon}(x_i | U_{\epsilon} = 0) &= p(x_i) \mathbb{E}[f | X_i = x_i] \\ p_{\epsilon}(x_i | U_{\epsilon} = 1) &= p(x_i) \left( \frac{1}{1 - \epsilon} - \frac{\epsilon}{1 - \epsilon} \mathbb{E}[f | X_i = x_i] \right) \end{aligned}$$

Thus, similarly,

$$\left. \frac{\partial}{\partial \epsilon} I_{\Phi}(U_{\epsilon}; X_i) \right|_{\epsilon=0} = H_{\Phi}(\mathbb{E}[f | X_i]).$$

Putting these together we obtain the desired equation.

## C Proof of Theorem 15

The equality  $\mathfrak{R}_{\Phi_{\alpha}}(X_{[k]}) = \mathfrak{R}_{\varphi_{\alpha}}(X_{[k]})$  for  $\alpha = 2$  is clear since  $\Phi_2 = \varphi_2$ . So we assume that  $\alpha \in (1, 2)$ . Moreover, the inclusion  $\mathfrak{R}_{\Phi_{\alpha}}(X_{[k]}) \subseteq \mathfrak{R}_{\varphi_{\alpha}}(X_{[k]})$  is immediate once we note that

$$\Phi_{\alpha}(t) = c_1(\varphi_{\alpha}(1 + t) + \varphi_{\alpha}(1 - t)) - c_2,$$

for some positive constants  $c_1, c_2$ . Then for any function  $f_{X_{[k]}}$ , we have

$$H_{\Phi_{\alpha}}(f) = c_1(H_{\varphi_{\alpha}}(1 + f) + H_{\varphi_{\alpha}}(1 - f)).$$

Writing the above equation for  $f$ , and  $\mathbb{E}[f | X_i]$  for  $i \in [k]$ , we obtain  $\mathfrak{R}_{\varphi_{\alpha}}(X_{[k]}) \subseteq \mathfrak{R}_{\Phi_{\alpha}}(X_{[k]})$ . So we need to prove the inclusion in the other direction.

Let  $\lambda_{[k]} \in \mathfrak{R}_{\Phi_{\alpha}}(X_{[k]})$ . We will show that  $\lambda_{[k]} \in \mathfrak{R}_{\varphi_{\alpha}}(X_{[k]})$ . Let  $f \geq 0$  be some non-negative function with  $m = \mathbb{E}f$ . Define

$$g_{\epsilon} = \epsilon f - 1.$$

Then for sufficiently small  $\epsilon \geq 0$  the range of  $g_{\epsilon}$  is inside the domain of  $\Phi_{\alpha}$ . For any such  $\epsilon$  we have

$$H_{\Phi_{\alpha}}(g_{\epsilon}) \geq \sum_{i=1}^k \lambda_i H_{\Phi_{\alpha}}(\mathbb{E}[g_{\epsilon} | X_i]). \quad (57)$$



We compute

$$\begin{aligned}
\Phi_\alpha(g_\epsilon) - \Phi_\alpha(\mathbb{E}g_\epsilon) &= \frac{1}{2^\alpha - 2} \left( (1 + g_\epsilon)^\alpha - (1 + \mathbb{E}g_\epsilon)^\alpha + (1 - g_\epsilon)^\alpha - (1 - \mathbb{E}g_\epsilon)^\alpha \right) \\
&= \frac{1}{2^\alpha - 2} \left( (\epsilon f)^\alpha - (\epsilon m)^\alpha + (2 - \epsilon f)^\alpha - (2 - \epsilon m)^\alpha \right) \\
&= \frac{1}{2^\alpha - 2} \left( \epsilon^\alpha (f^\alpha - m^\alpha) + (2 - \epsilon f)^\alpha - (2 - \epsilon m)^\alpha \right).
\end{aligned}$$

Taking expectation from both sides we obtain

$$\begin{aligned}
H_{\Phi_\alpha}(g_\epsilon) &= \frac{1}{2^\alpha - 2} \left( \epsilon^\alpha (\mathbb{E}[f^\alpha] - m^\alpha) + \mathbb{E}[(2 - \epsilon f)^\alpha - (2 - \epsilon m)^\alpha] \right) \\
&= \frac{\epsilon^\alpha}{2^\alpha - 2} H_{\varphi_\alpha}(f) + O(\epsilon^2),
\end{aligned}$$

where the  $O(\epsilon^2)$  term is derived once we take the Taylor expansion of  $(2 - t)^\alpha$ . We similarly have

$$H_{\Phi_\alpha}(\mathbb{E}[g_\epsilon|X_i]) = \frac{\epsilon^\alpha}{2^\alpha - 2} H_{\varphi_\alpha}(\mathbb{E}[f|X_i]) + O(\epsilon^2).$$

Putting these in (57) and using the fact that  $1 < \alpha < 2$  we find that

$$H_{\varphi_\alpha}(f) \geq \sum_{i=1}^k \lambda_i H_{\varphi_\alpha}(\mathbb{E}[f|X_i]).$$

Therefore,  $\lambda_{[k]} \in \mathfrak{R}_{\varphi_\alpha}(X_{[k]})$ .

## D Proof of Theorem 26

Using Proposition 24 and the inequality  $\text{Var}[\mathbb{E}[f|X_i]] \geq \mathbb{E}[f\hat{f}_i]^2$ , the inclusion  $\mathfrak{S}(X_{[k]}) \subseteq \mathfrak{S}'(X_{[k]})$  is immediate. It suffices to show that  $\mathfrak{S}'(X_{[k]}) \subseteq \mathfrak{S}(X_{[k]})$ .

Let  $\lambda_{[k]} \in \mathfrak{S}'(X_{[k]})$  and let  $f_i(X_i)$ ,  $i = 1, \dots, k$ , be arbitrary functions with zero mean. According to the characterization of Theorem 25 of the MC ribbon we need to show that

$$\text{Var}[f_1 + \dots + f_k] \leq \sum_{i=1}^k \frac{1}{\lambda_i} \text{Var}[f_i].$$

Let  $c_i = \sqrt{\text{Var}[f_i]}$  and define  $\hat{f}_i = f/c_i$ . Also define  $m_{ij} = \mathbb{E}[\hat{f}_i \hat{f}_j]$ . Observe that

$$\text{Var}[f] = \sum_{i,j=1}^k c_i c_j m_{ij},$$

and

$$\mathbb{E}[f\hat{f}_i]^2 = \left( \sum_{j=1}^k c_j m_{ij} \right)^2.$$

Now since  $\lambda_{[k]} \in \mathfrak{S}'(X_{[k]})$  we have

$$\sum_{i,j=1}^k c_i c_j m_{ij} \geq \sum_{i=1}^k \lambda_i \left( \sum_{j=1}^k c_j m_{ij} \right)^2.$$

Indeed, this inequality must hold for all choices of  $c_i$ 's.

Let  $M$  be a matrix whose  $(i, j)$ -th entry is  $m_{ij}$ . Also let  $\Lambda$  be a diagonal matrix with diagonal entries equal to  $\lambda_1, \dots, \lambda_k$ . Then, the above inequality, for all choices of  $c_i$ 's, is equivalent to  $M \geq M\Lambda M$ , which itself is equivalent to  $M^{-1} \geq \Lambda$ . Next, since  $t \mapsto -t^{-1}$  is operator monotone, it is also equivalent to  $\Lambda^{-1} \geq M$ . Then by simple calculation  $\Lambda^{-1} \geq M$  means that for all  $c_1, \dots, c_k$  we have

$$\text{Var}[c_1 \hat{f}_1 + \dots + c_k \hat{f}_k] \leq \sum_{i=1}^k \frac{1}{\lambda_i} c_i^2 = \sum_{i=1}^k \frac{1}{\lambda_i} \text{Var}[c_i \hat{f}_i].$$

This is what we wanted to show.

## E Proof of Proposition 30

We use Theorem 25 to prove this proposition. Any zero-mean function of  $X_i$ ,  $i = 1, 2$ , is of the form  $f_i = a_i g_i$  for some constants  $a_1, a_2$ . Moreover, the space of zero-mean functions of  $X_3$  is two-dimensional. Since we assume that  $\mathbb{E}[g_1|X_3]$  and  $\mathbb{E}[g_2|X_3]$  are linearly independent, this space is spanned by these two functions. That is, any zero-mean function  $f_3(X_3)$  can be expressed as

$$f_3 = a_3 \mathbb{E}[g_1|X_3] + a_4 \mathbb{E}[g_2|X_3],$$

for some constants  $a_3$  and  $a_4$ . Then using equations (45)-(48) we have

$$\begin{aligned} \text{Var}[f_1] &= a_1^2, \\ \text{Var}[f_2] &= a_2^2, \\ \text{Var}[f] &= a_3^2 \rho_{13}^2 + a_4^2 \rho_{23}^2 + 2a_3 a_4 r_{12 \rightarrow 3}, \\ \mathbb{E}[f_1 f_2] &= a_1 a_2 \rho_{12}, \\ \mathbb{E}[f_1 f_3] &= a_1 (a_3 \rho_{13}^2 + a_4 r_{1,2 \rightarrow 3}), \\ \mathbb{E}[f_2 f_3] &= a_2 (a_3 r_{1,2 \rightarrow 3} + a_4 \rho_{2,3}^2). \end{aligned}$$

Hence,

$$\begin{aligned} \text{Var}\left[\sum_{i=1}^k f_i\right] &= a_1^2 + a_2^2 + a_3^2 \rho_{1,3}^2 + a_4^2 \rho_{2,3}^2 + 2a_3 a_4 r_{1,2 \rightarrow 3} \\ &\quad + 2a_1 a_2 \rho_{1,2} + 2a_1 (a_3 \rho_{1,3}^2 + a_4 r_{1,2 \rightarrow 3}) + 2a_2 (a_3 r_{1,2 \rightarrow 3} + a_4 \rho_{2,3}^2). \end{aligned}$$

On the other hand,  $\mathfrak{S}(X_1, X_2, X_3)$  is the set of triples  $(\lambda_1, \lambda_2, \lambda_3)$  such that

$$\text{Var}\left[\sum_{i=1}^k f_i\right] \leq \sum_{i=1}^3 \frac{1}{\lambda_i} \text{Var}[f_i],$$

for all choices of  $a_1, \dots, a_4$ . Using the previous equations, the above inequality is a quadratic form in terms of  $a_1, \dots, a_4$ . Indeed, letting  $\mathbf{v} = [a_1, a_2, a_3, a_4]$ , the above inequality is equivalent to  $\mathbf{v} \Delta \mathbf{v}^t \geq 0$  for all  $\mathbf{v}$ , where

$$\Delta = \begin{bmatrix} \frac{1}{\lambda_1} - 1 & -\rho_{12} & -\rho_{13}^2 & -r_{12 \rightarrow 3} \\ -\rho_{12} & \frac{1}{\lambda_2} - 1 & -r_{12 \rightarrow 3} & -\rho_{23}^2 \\ -\rho_{13}^2 & -r_{1,2 \rightarrow 3} & \rho_{13}^2 \left(\frac{1}{\lambda_3} - 1\right) & r_{1,2 \rightarrow 3} \left(\frac{1}{\lambda_3} - 1\right) \\ -r_{1,2 \rightarrow 3} & -\rho_{23}^2 & r_{1,2 \rightarrow 3} \left(\frac{1}{\lambda_3} - 1\right) & \rho_{23}^2 \left(\frac{1}{\lambda_3} - 1\right) \end{bmatrix}. \quad (58)$$

In other words,  $\mathfrak{S}(X_1, X_2, X_3)$  is the set of  $(\lambda_1, \lambda_2, \lambda_3)$  for which  $\Delta$  is positive semi-definite.

Observe that  $\Delta$  can be written in the block form:

$$\Delta = \begin{bmatrix} A & -B \\ -B & (\frac{1}{\lambda_3} - 1)B \end{bmatrix},$$

where  $A$  and  $B$  are  $2 \times 2$  matrices. Then using [21, p.14],  $\Delta$  is positive semi-definite if and only if

$$A - B(\frac{1}{\lambda_3} - 1)^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} - 1 - \rho_{1,3}^2(\frac{1}{\lambda_3} - 1)^{-1} & -\rho_{1,2} - r_{1,2 \rightarrow 3}(\frac{1}{\lambda_3} - 1)^{-1} \\ -\rho_{1,2} - r_{1,2 \rightarrow 3}(\frac{1}{\lambda_3} - 1)^{-1} & \frac{1}{\lambda_2} - 1 - \rho_{2,3}^2(\frac{1}{\lambda_3} - 1)^{-1} \end{bmatrix},$$

is positive semi-definite. This is equivalent with the conditions given in the statement of the theorem.

## F Proof of Theorem 31

When  $X_i$ 's are binary, then any function of  $X_i$  with zero mean is of the form  $f_i(X_i) = a_i(X_i - \mathbb{E}[X_i])$  for some constant  $a_i$ . Using this fact it is easy to see that

$$\text{Var}\left[\sum_i f_i\right] \leq \sum_i \frac{1}{\lambda_i} \text{Var}[f_i], \quad (59)$$

holds for all choices of  $a_i$ 's if and only if  $R \leq \Lambda^{-1}$ .

Let us turn to the proof for Gaussian variables. Our proof is an extension of the proof of Lancaster [15] to the multivariate case.

Observe that scaling of and adding a constant to the variables  $X_i$  would not change the MC ribbon. Hence, without loss of generality we assume that  $\mathbb{E}[X_i] = 0$ ,  $\text{Var}[X_i] = 1$ , and  $\mathbb{E}[X_i X_j] = R_{ij}$ .

The Hermite-Tchebycheff polynomials are defined as follows:

$$\psi_\ell(x) = (-1)^\ell e^{x^2} \frac{d^\ell}{dx^\ell} e^{-x^2}, \quad \ell \geq 0.$$

The following facts are known about these polynomials [15]:

- (i)  $\psi_0(x) = 1$  is a constant function, and  $\psi_i(x)$  and  $\psi_j(x)$  are orthonormal with respect to standard normal distribution, i.e.,

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_i(x) \psi_j(x) e^{-\frac{x^2}{2}} dx = \delta_{ij}. \quad \forall i, j.$$

- (ii) If  $X$  is a normal random variable, any function of  $X$  denoted by  $f(X)$  that has finite variance can be approximated as follows: for any  $\epsilon > 0$ , there is a sequence  $\{a_\ell | \ell \geq 0\}$  such that  $\sum_\ell a_\ell^2$  is convergent, and for

$$\hat{f}(x) = \sum_{\ell=0}^{\infty} a_\ell \psi_\ell(x),$$

we have

$$\mathbb{E}\left[|f(X) - \hat{f}(X)|^2\right] \leq \epsilon.$$

Furthermore, if  $\mathbb{E}[f(X)] = 0$ , we may take  $a_0 = 0$ .

(iii) If  $X$  and  $Y$  are unit variance, jointly Gaussian random variables with correlation coefficient  $\rho$ , then

$$\mathbb{E}[\psi_\ell(X)\psi_{\ell'}(Y)] = \delta_{\ell\ell'}\rho^\ell.$$

Fix  $(\lambda_1, \dots, \lambda_k) \in [0, 1]^k$ . To verify the validity of (59) take some arbitrary zero-mean functions  $f_i(X_i)$ ,  $i = 1, \dots, k$ , with finite variance. Fix some  $\epsilon > 0$ , and using property (ii) explained above let

$$\hat{f}_i(x_i) = \sum_{\ell=1}^{\infty} a_{i\ell} \psi_\ell(x_i), \quad (60)$$

be such that

$$\mathbb{E}\left[|f_i(X_i) - \hat{f}_i(X_i)|^2\right] \leq \epsilon, \quad \forall i. \quad (61)$$

Then, by the Cauchy-Schwarz inequality we have

$$\mathbb{E}\left[\left|\sum_{i=1}^k f_i(X_i) - \sum_{i=1}^k \hat{f}_i(X_i)\right|^2\right] \leq k\epsilon. \quad (62)$$

Then, it is not hard to verify that (61) implies

$$\left|\text{Var}[\hat{f}_i(X_i)] - \text{Var}[f_i(X_i)]\right| = O(\sqrt{\epsilon}),$$

and similarly (62) implies

$$\left|\text{Var}\left[\sum_{i=1}^k \hat{f}_i(X_i)\right] - \text{Var}\left[\sum_{i=1}^k f_i(X_i)\right]\right| = O(\sqrt{\epsilon}).$$

Therefore, it suffices to verify (59) for functions of the form (60).

Using properties (i) and (iii) we have

$$\text{Var}[\hat{f}_i(X_i)] = \sum_{\ell=1}^{\infty} a_{i\ell}^2,$$

and

$$\begin{aligned} \text{Var}\left[\sum_{i=1}^k \hat{f}_i(X_i)\right] &= \sum_{i_1, i_2=1}^k \mathbb{E}[\hat{f}_{i_1}(X_{i_1})\hat{f}_{i_2}(X_{i_2})] \\ &= \sum_{i_1, i_2=1}^k \sum_{\ell_1, \ell_2=1}^{\infty} a_{i_1\ell_1} a_{i_2\ell_2} \mathbb{E}[\psi_{\ell_1}(X_{i_1})\psi_{\ell_2}(X_{i_2})] \\ &= \sum_{i_1, i_2=1}^k \sum_{\ell=1}^{\infty} a_{i_1\ell} a_{i_2\ell} R_{i_1, i_2}^\ell. \end{aligned}$$

Thus, we are interested in the set of  $k$ -tuples  $(\lambda_1, \dots, \lambda_k)$  such that for all  $a_{i\ell}$ 's we have

$$\sum_{i_1, i_2=1}^k \sum_{\ell=1}^{\infty} a_{i_1\ell} a_{i_2\ell} R_{i_1, i_2}^\ell \leq \sum_{i=1}^k \frac{1}{\lambda_i} \sum_{\ell=1}^{\infty} a_{i\ell}^2.$$

This holds if and only if for any  $\ell \in \mathbb{N}$  and for any  $a_{i\ell}$ 's we have

$$\sum_{i_1, i_2=1}^k a_{i_1\ell} a_{i_2\ell} R_{i_1, i_2}^\ell \leq \sum_{i=1}^k \frac{1}{\lambda_i} a_{i\ell}^2.$$

This can be expressed in matrix form as

$$R^{\circ\ell} \leq \Lambda^{-1} \quad \forall \ell \geq 1, \quad (63)$$

where  $R^{\circ\ell}$  is the Hadamard product (entry-wise product) of  $R$  with itself  $\ell$  times. For  $\ell = 1$ , we have the condition

$$R \leq \Lambda^{-1}. \quad (64)$$

Now, we claim that (64) implies (63) for any  $\ell \geq 2$ . To prove this, note that  $R \leq \Lambda^{-1}$  means that  $\Lambda^{-1} - R \geq 0$  is positive semi-definite. Moreover,  $R$  is a correlation matrix, so it is positive semi-definite. Since the Hadamard product of two positive semi-definite matrix is positive semi-definite,  $R^{\circ(\ell-1)}$  is positive semi-definite as well. Similarly,  $R^{\circ(\ell-1)} \circ (\Lambda^{-1} - R)$  is positive semi-definite, *i.e.*,  $R^{\circ(\ell-1)} \circ \Lambda^{-1} \geq R^{\circ\ell}$ . Now the point is that the diagonal entries of  $R$  are all one, and  $\Lambda$  is diagonal. Therefore,  $R^{\circ(\ell-1)} \circ \Lambda^{-1} = \Lambda^{-1}$ . This completes the proof.

## G Proofs of Theorem 33

**Tensorization:** Assuming that  $X_{[k]}$  and  $Y_{[k]}$  are independent, we would like to show that

$$\tilde{\mathfrak{S}}(X_1 Y_1, \dots, X_k Y_k) = \tilde{\mathfrak{S}}(X_1, \dots, X_k) \cap \mathfrak{S}(Y_1, \dots, Y_k)$$

We clearly have

$$\tilde{\mathfrak{S}}(X_1 Y_1, \dots, X_k Y_k) \subseteq \tilde{\mathfrak{S}}(X_1, \dots, X_k) \cap \mathfrak{S}(Y_1, \dots, Y_k)$$

since we can restrict to functions  $f_i(X_i, Y_i)$  to depend only on one of  $X_i, Y_i$ . To prove the inclusion in the other direction, take some  $(\lambda_1, \dots, \lambda_k) \in \tilde{\mathfrak{S}}(X_1, \dots, X_k) \cap \tilde{\mathfrak{S}}(Y_1, \dots, Y_k)$ . For arbitrary functions  $f_i(X_i, Y_i)$ ,  $i = 1, \dots, k$ , we compute

$$\text{Var}_{X_{[k]} Y_{[k]}} \left[ \sum_{i=1}^k f_i \right] = \text{Var} \left[ \mathbb{E} \left[ \sum_{i=1}^k f_i \middle| X_{[k]} \right] \right] + \text{Var} \left[ \sum_{i=1}^k f_i \middle| X_{[k]} \right] \quad (65)$$

$$= \text{Var} \left[ \sum_{i=1}^k \mathbb{E} [f_i | X_i] \right] + \text{Var} \left[ \sum_{i=1}^k f_i \middle| X_{[k]} \right] \quad (66)$$

$$\geq \sum_{i=1}^k \lambda_i \text{Var} [\mathbb{E} [f_i | X_i]] + \sum_{i=1}^k \lambda_i \text{Var} [f_i | X_{[k]}] \quad (67)$$

$$= \sum_{i=1}^k \lambda_i \text{Var} [\mathbb{E} [f_i | X_i]] + \sum_{i=1}^k \lambda_i \text{Var} [f_i | X_i] \quad (68)$$

$$= \sum_{i=1}^k \lambda_i \text{Var} [f_i]. \quad (69)$$

Here equations (65) and (69) follow from the law of total variance. Equations (66) and (68) follow from the independence of  $X_{[k]}$  and  $Y_{[k]}$ . Finally, (67) holds since  $(\lambda_1, \dots, \lambda_k)$  is in both  $\tilde{\mathfrak{S}}(X_{[k]})$  and  $\tilde{\mathfrak{S}}(Y_{[k]})$ .

**Monotonicity:** Let  $(\lambda_1, \dots, \lambda_k) \in \tilde{\mathfrak{S}}(X_1, \dots, X_k)$  and let  $f_i(Y_i)$ ,  $i = 1, \dots, k$ , be arbitrary functions. We need to show that

$$\text{Var}\left[\sum_{i=1}^k f_i\right] \geq \sum_{i=1}^k \lambda_i \text{Var}[f_i]. \quad (70)$$

For functions  $\mathbb{E}[f_i|X_i]$ ,  $i = 1, \dots, k$ , we have

$$\text{Var}\left[\mathbb{E}\left[\sum_{i=1}^k f_i \middle| X_{[k]}\right]\right] = \text{Var}\left[\sum_{i=1}^k \mathbb{E}[f_i|X_i]\right] \geq \sum_{i=1}^k \lambda_i \text{Var}[\mathbb{E}[f_i|X_i]].$$

Moreover, since  $Y_i$ 's are independent conditioned on  $X_{[k]}$  we have

$$\text{Var}\left[\sum_{i=1}^k f_i \middle| X_{[k]}\right] = \sum_{i=1}^k \text{Var}[f_i|X_{[k]}] = \sum_{i=1}^k \text{Var}[f_i|X_i] \geq \sum_{i=1}^k \lambda_i \text{Var}[f_i|X_i].$$

Summing up the above two inequalities and using the law of total variance, we obtain (70).

## H Proof of Theorem 35

If  $(f_1, \dots, f_k)$  with the conditions in the theorem exists, we clearly have  $\tilde{\mathfrak{S}}(X_{[k]}) = \{(0, \dots, 0)\}$ .

Now suppose that  $\tilde{\mathfrak{S}}(X_{[k]}) = \{(0, \dots, 0)\}$ . Then for any  $\epsilon > 0$  there are functions  $f_i^{(\epsilon)}(X_i)$ ,  $i = 1, \dots, k$ , such that

$$\text{Var}\left[f_1^{(\epsilon)} + \dots + f_k^{(\epsilon)}\right] \leq \epsilon \sum_{i=1}^k \text{Var}\left[f_i^{(\epsilon)}\right],$$

and

$$\sum_{i=1}^k \text{Var}\left[f_i^{(\epsilon)}\right] = 1.$$

Then by a compactness argument, there are limiting functions  $\hat{f}_i(X_i)$ ,  $i = 1, \dots, k$ , such that

$$\sum_{i=1}^k \text{Var}[\hat{f}_i] = 1,$$

and

$$\text{Var}[\hat{f}_1 + \dots + \hat{f}_k] = 0.$$

Since  $\mathbb{E}[\hat{f}_i] = 0$  for all  $i$ , the latter equation means that  $\hat{f}_1 + \dots + \hat{f}_k = 0$ . Furthermore,  $(\hat{f}_1, \dots, \hat{f}_k)$  is non-zero because of  $\sum_{i=1}^k \text{Var}[\hat{f}_i] = 1$ . We are done.